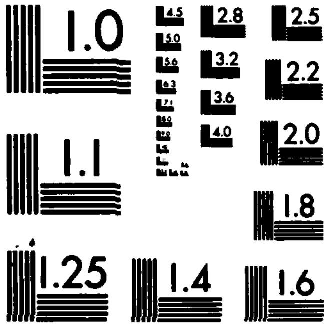
AD-A140 468 LOW-LEVEL STRATUS PREDICTION USING BINARY STATISTICAL REGRESSION: A PROGR..(U) NAVAL POSTGRADUATE SCHOOL MONTEREY CA D P GAVER ET AL. DEC 83 NPS55-83-034 1/1

UNCLASSIFIED F/G 12/1 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

(2)

NPS55-83-034

NAVAL POSTGRADUATE SCHOOL

Monterey, California

LOW-LEVEL STRATUS PREDICTION USING BINARY STATISTICAL REGRESSION: A PROGRESS REPORT USING MOFFETT FIELD DATA

by

Donald P. Gaver

Patricia A. Jacobs

December 1983

Prepared for:
Chief of Naval Research
Arlington, VA 22217

NAVAL POSTGRADUATE SCHOOL
Monterey, California

Commodore R. H. Shumaker
Superintendent

David A. Schrady
Provost

This work was supported by the Naval Environmental Prediction Research Facility, and by the Probability and Statistics Program of the Office of Naval Research.

D. P. Gaver
Professor
Department of Operations Research

Patricia A. Jacobs
Associate Professor
Department of Operations Research

Reviewed by:

Alan R. Washburn, Chairman
Department of Operations Research

Released by:

Kneale T. Marshall
Dean of Information and Policy Sciences

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE — READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
NPS55-83-034

2. GOVT ACCESSION NO.
AD-A140468

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
LOW-LEVEL STRATUS PREDICTION USING BINARY STATISTICAL REGRESSION: A PROGRESS REPORT USING MOFFETT FIELD DATA

5. TYPE OF REPORT & PERIOD COVERED
Technical Report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Donald P. Gaver
Patricia A. Jacobs

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Naval Postgraduate School
Monterey, CA 93943

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
61153N; RR014-05-OE
N0001484WR24011

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
Arlington, VA 22217

12. REPORT DATE
December 1983

13. NUMBER OF PAGES
77

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Single station prediction of stratus; logistic model; dewpoint depression; robust estimation; shrinkage

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Various statistical models and techniques were employed to forecast the existence of low-level stratus conditions. They are illustrated for data at a single-station (Moffett Field, Sunnyvale, California) using single-station surface meteorological measurements only as explanatory variables. A preliminary exploratory data analysis shows that low (high) dew point depression is associated with the existence (non-existence) of low-level stratus at Moffett Field. Procedures for and results of various methods of

fitting logistic models to the data are described. The fitted models were used to forecast stratus on reserved data sets (cross-validation). Results of the cross-validation are given.

Abstract

Various statistical models and techniques were employed to forecast the existence of low-level stratus conditions. They are illustrated for data at a single station (Moffett Field, Sunnyvale, California) using single-station surface meteorological measurements only as explanatory variables. A preliminary exploratory data analysis shows that low (high) dew point depression is associated with the existence (non-existence) of low-level stratus at Moffett Field. Procedures for and results of various methods of fitting logistic models to the data are described. The fitted models were used to forecast stratus on reserved data sets (cross-validation). Results of the cross-validation are given.

LOW-LEVEL STRATUS PREDICTION USING BINARY STATISTICAL REGRESSION: A PROGRESS REPORT, USING MOFFETT FIELD DATA

Donald P. Gaver   Patricia A. Jacobs

Operations Research Department
Naval Postgraduate School

0. Executive Summary

In this paper various statistical models and techniques are employed to forecast the existence of low-level stratus conditions. They are illustrated for data at an airport (Moffett Field, Sunnyvale, California).

In Section 2 the data set is described and the results of a preliminary exploratory data analysis are given. These suggest that dew point depression should be predictive of the existence of stratus. Generally, low (high) dew point depression is associated with the existence (non-existence) of stratus. This association is also made evident by a spectral analysis of hourly stratus levels and dewpoint depression described in Appendix F.

The remainder of this paper describes procedures for and results of, fitting logistic models to the data described. Validation of the models are addressed as well. The basic logistic model is

$$P\{Y = 1 | \text{explanatory variable } \underline{x}\} = \frac{\exp\{\underline{x}\ \underline{\beta}\}}{1 + \exp\{\underline{x}\ \underline{\beta}\}}$$

where $\underline{x}$ is a p-vector (row) of explanatory variables and $\underline{\beta}$ is a p-vector (column) of coefficients to be determined.

Appendix E suggests several mathematical justifications for use of the logistic regression model.

We have used various methods to fit logistic models for use as predictors on reserved data sets (cross-validation). Our cross-validation experiences are reported in Appendices A through D. Appendix G contains the asymptotic distribution of a threat score, which is one of the statistics we use to compare procedures.

Appendix A reports on use of the stepwise logistic regression procedure of the BMDP computer package. The procedure chooses variables to be used in the regression from a menu of variables given to it. The BMDP fits are then used to predict the occurrence of stratus for independent data, i.e. from different years. We find that the stepwise feature must be used with caution; it tends to overfit, including variables which greatly increase the standard error of the variables first included in the regression. Such overfitting degrades the predictive powers of the model.

Copas (1983) points out that a regression model, fit by maximum likelihood (or least squares) to one set of data, and then used for prediction on another set of data, nearly always fits or predicts the new set of data less well than it does the original set. This phenomenon of shrinkage can become more pronounced if the original regression model is fit using a stepwise procedure, which tends to overfit. Appendix B describes and investigates a procedure suggested by Copas to compensate for shrinkage on both regressions fit with, and without, stepwise

procedures. In our application, particularly when predicting changes from stratus to no stratus, the shrinkage procedure appears to help. However, it appears to do less well in predicting changes from no stratus to stratus.

In Appendix C robust estimation procedures for logistic regression are described and carried out on the Moffett Field data. These procedures are less vulnerable than maximum likelihood estimates to a few outlying data points which may not agree with the model. For the particular cross validations performed, predictions using models fit with robust procedures were no better than predictions made with the models fit with maximum likelihood. The models obtained are, however, systematically different from their classical counterparts.

In Appendix D we investigate the predictive use of logistic regression models that are progressively updated to emphasize recent data. The suggestion is that models fit with data which are closer in time to the dates on which forecasts are to be made may be more relevant, owing to changing conditions not represented in the model, than a model which is fit with data of several previous years. We found that models with updating often did at least as well as models without an updating feature.

In summary, we have found that ℓn(dew point depression + 1) appears to be a consistently useful predictor of the occurrence of stratus. Low (high) dew point depression is associated with stratus (no stratus). There is no one procedure or model, among those tried to date, that appears a clear winner. If, for

example, one procedure does well in predicting changes from no stratus to stratus, it will often do less well in predicting changes from stratus to no stratus. We found that none of the procedures did as well predicting the occurrence of stratus in 1962 as it did in 1961. This suggests that perhaps 1962 is not described by the present models as well as is 1961, being intrinsically quite different from the previous years 1958-61. Models and methods that represent year-to-year differences will come under investigation in future.

Further work, with other models, and with data from other locations, will be undertaken to shed light on this important prediction problem.

1. Introduction and Overview

The purpose of this paper is to exhibit the use of statistical tools and procedures for forecasting the existence of low-level stratus conditions at an airport. The existence of low stratus (less than or equal to 1000 ft.) forces the use of different methods of traffic handling than is the case when higher stratus levels prevail. A low stratus condition tends to inhibit flight operations, so it is desirable to forecast its occurrence. Furthermore, it is of interest to forecast such conditions on a "single-station" basis, making use of meteorological measurements available only at the location--e.g. airport--in question, in case useful supplementary information is unavailable.

The forecasting approaches described here are statistical in nature, meaning that extensive data concerning the reported hourly stratus level at an airport (Moffett Field, Sunnyvale, CA), together with certain other meteorological measurements or parameters recorded and reported at that location, were used as raw material for the forecasts. These data were used to estimate the probability of low stratus during a daily period; the latter probability was estimated using a logistic regression model, a tool that has been found useful in biological and medical statistics, and that has been previously applied in meteorology; cf Brelsford and Jones (1967), Gilhausen (1979), Gabriel and Pun (1979). In a later section we present various derivations or justifications of such a model. Alternative models are also suggested, and the usefulness of these will be investigated in future work.

The usefulness of the logistic (or any other) model must be judged by its performance. We have chosen to proceed by (i) fitting a model to data for certain specific years (1958-1960), and then (ii) comparing the model predictions to actual occurrences for a completely different period (1961, 1962). Such a procedure is termed cross validation; see Mosteller and Tukey (1977) for good general discussion and references. The results of our cross validation are reported subsequently. Another interesting and possibly useful approach is to construct and test an adaptive, automatically up-dating forecasting model with characteristics similar to "exponential smoothing" or "Kalman filtering." Results of some simple updating procedures for forecasting will also be reported.

Successful forecasting with the aid of a model requires that the data inputs be relatively "clean," or in basic conformity with the model. Occasionally occurring data points that are out of line for any reason, called outliers, or influential values, can radically change the values of the model parameters obtained from statistical fitting principles such as least squares (not used for fitting our logistic model) or maximum likelihood (which is used). To check for such maverick, possibly detrimentally influential, values it is possible to proceed in several ways. One is to successively remove each data point (actually a vector of response and explanatory variables) and re-fit the model, watching for radical changes in fitted model parameters. This method has been programmed (in APL, on the NPS IBM 3033 system) and exercised; its defect is that at present just one data point

is removed at a time, so if several points are mavericks this fact may be overlooked. Clever ways of automatically diminishing the effects of maverick points have been discussed by Pregibon (1982); exploration of the applicability of such ideas to the present stratus prediction problem is currently underway. The methods and some results are reported here.

Another approach to the identification of maverick data, and to the possible discovery of an appropriate model, is by computer graphics. We have initiated the examination of the low-stratus data on a pioneering graphics facility at Stanford Linear Accelerator (SLAC); see an article in Science, Kolata (1982), for general description. The SLAC system allows an analyst views of various three-dimensional space projections of multidimensional data-clouds. Such examination helps to reveal the association between certain explanatory ("independent") variables and the response ("dependent variable") of interest. For example, examination of our stratus data indicated that changes in the explanatory variable dewpoint depression tended to be reflected in changes of response, i.e. low stratus level probability. This association has physical basis, and dewpoint depression had actually been included in earlier exploratory logistic fits at the suggestion of W. Sweet of NEPRF; its incorporation into the model considerably improves predictive performance.

2. The Basic Data Set

The statistical methods used in this study were applied to data furnished by W. Sweet of NEPRF, to whom we are grateful. In summary, these data consist of reported hourly determinations of:

(i) stratus level, reported to be at discrete levels of 100 ft. separation; possible recorded levels are $k \times 100$ ft., $k = 1,2,...,9,10,...,$"999" (no visible stratus).

(ii) east-west wind velocity, V_x, at surface, in miles per hour;

(iii) north-south wind velocity, V_y, at surface, in miles per hour;

(iv) temperature, at surface, in degrees F;

(v) dewpoint, at surface, in degrees F;

all at Moffett Field, California, for the months of July, August, and September of the years 1958-1962; later data are also available, and remain to be analyzed. Although other measurements, e.g. of pressure, are in principle available, they were not utilized in the present analysis. Nor were measurements from neighboring locations in the San Francisco Bay area.

2.1. The Forecasting Exercise Data Set

The raw data described above were adopted to the forecasting exercises as follows:

(a) Forecasts are made of the existence of stratus level less than 1000 ft. ($\leq$ 900 ft.) on any hour between 10:00 pm (2200) on day t, and 6:00 am (0600) on day $t + 1$. If hourly-reported stratus level ever fell to a level $\leq$ 900 ft. during such a period beginning on day t, it is agreed to say that stratus existed on day t; otherwise

that no stratus existed on day t. Denote by the binary indicator variable y_t the existence (non-existence) of stratus on day t according to the above definition. Thus

$$y_t = \begin{cases} 1 & \text{if stratus exists on day } t, \\ 0 & \text{if no stratus exists on day } t. \end{cases}$$

Call y_t the <u>response</u> (or dependent variable) when forecasting for day t. Note that the observed values of response on previous days $(y_{t-1}, y_{t-2}, \ldots)$ are available as assistance when forecasting for day t. The above definition of meaningful stratus agrees with instrument/no instrument landing rules at airports, and is thus of operational significance.

Candidate <u>explanatory</u> (independent) variables are these:

(b) wind velocities at 6:00 pm (1800) on day t, items (ii) and (iii) above;

(c) temperature (T_t) and dewpoint (D_t) at surface at 6:00 pm on day t;

(d) dewpoint depression, $\bar{\Delta}_t = T_t - D_t$ at 6:00 pm on day t;

(e) hours of stratus (H_{t-1}) between 2200 on the previous day $t-1$ and 0600 on the current day t;

(f) existence/non-existence of stratus $(y_{t-1}, y_{t-2}, \ldots)$ on previous days.

Let NS_t denote the number of consecutive days of stratus in a run of stratus days that includes day $t-1$, the day on which the prediction is made. NNS_t is the number of consecutive days of no-stratus in a run of no-stratus days that includes day $t-1$.

Note that because of the way in which the response y_t is defined, it is legitimate and of interest to forecast y_t in terms of T_t, D_t, $\bar{\Delta}_t$, $V_x(t)$, etc. These latter quantities are all available at 6:00 pm for forecasts applying later, i.e. from 10:00 pm to 6:00 am on the following day. Of course many other functions of the hourly observations are candidates for explanatory variable status.

3. Preliminary Analysis

Before proceeding to the fitting of specific models, a subset of the data has been examined in terms of simple summaries. Since the objective is to forecast, we have divided (conditioned) the data for the years 1958, 1959, 1960 into four groups:

Group 00: observations such that $y_{t-1} = 0$, $y_t = 0$,

Group 01: observations such that $y_{t-1} = 0$, $y_t = 1$,

Group 10: observations such that $y_{t-1} = 1$, $y_t = 0$,

Group 11: observations such that $y_{t-1} = 1$, $y_t = 1$,

and have then computed summaries of the observed distributions of certain candidate explanatory variables. The argument is that a noticeable separation of such distributions when predicting y_t from the particular explanatory data suggests that the variable in question may be useful in forecasting.

Note that we have explicitly used the known stratus state of the system at $t-1$ as one important variable, wishing to make full use of persistence, and to improve upon it. We are especially interested in the power of explanatory variables and their combinations to correctly forecast <u>changes</u> in stratus conditions, e.g. from $y_{t-1} = 0$ (no stratus on day $t-1$) to $y_t = 1$ (stratus on day t). Simple persistence forecasting, which predicts $y_t = y_{t-1}$ will never identify prospective changes.

Computer graphic analysis carried out at SLAC, plus physical insight, suggest that dewpoint depression, $\bar{\Delta}_t$, should be an effective explanatory variable. Another useful variable

seems to be the hours of stratus observed on day t-1, denoted by H_{t-1}. There are limitless other plausible explanatory variables, as well as combinations and re-expressions (transformations) of the latter, but here we look at only two. One systematic way of uncovering predictive combinations of explanatory variables is by use of some form of principle component or factor analysis; such work is not reported here. It seems possible that a <u>robust</u> principle component analysis may be informative (see Gnanadesikan (1977), or Campbell (1982)), for the existence of groups of maverick-like data have been reported in the overall data base. Clustering procedures may also be of value.

Tables 1 and 2 give a few useful summaries of the behavior of the candidate explanatory variables $\bar{\Delta}_t$ and H_{t-1}; these have been developed for the years 1958, 1959, 1960. The figures in parentheses are natural logarithms of their counterparts. The log transformation is suggested to symmetrize the sample distribution (histogram or Tukey stem-leaf plot), which often tends to appear positively skewed for the above data. The medians and quartiles are used instead of the ordinary means and standard deviations because of the possible non-robust/resistant properties of the latter traditional measures.

We can draw the following conclusions from Table 1:

(a) corresponding summary figures for dewpoint depression ($\underline{Q}$,M,$\bar{Q}$) are rather stable from year to year.

(b) dewpoint depression (or its log) should have prognostic power: roughly speaking,

TABLE 1

Observed Distribution of Dew Point Depression ($\bar{\Delta}_t$)

Year		Lower Quartile ($\underline{Q}$)	Median (M)	Upper Quartile ($\bar{Q}$)
1958;	1 → 0:	9(2.2)	9(2.2)	10(2.3)
	1 → 1:	6(1.79)	7(1.95)	9(2.2)
	0 → 1:	6(1.79)	8(2.08)	8(2.08)
	0 → 0:	10(2.3)	13(2.56)	17(2.83)
1959;	1 → 0:	8(2.08)	9(2.2)	11(2.4)
	1 → 1:	6(1.79)	7(1.95)	9(2.2)
	0 → 1:	7(1.95)	9(2.2)	10(2.3)
	0 → 0:	10(2.3)	14(2.64)	16(2.77)
1960;	1 → 0:	7(1.95)	10(2.3)	13(2.56)
	1 → 1:	6(1.79)	8(2.08)	9(2.2)
	0 → 1:	7(1.95)	8(2.08)	11(2.4)
	0 → 0:	10(2.3)	13(2.56)	16(2.77)

(b-1) if stratus is present at time (day) $t-1$, and if $\bar{\Delta}_t$ is <u>relatively high</u> (9 or above), a <u>change to no stratus</u> is indicated, while if $\bar{\Delta}_t$ is <u>relatively low</u> (below 9) the <u>stratus condition tends to continue</u>; on the other hand

(b-2) if no stratus is present at time (day) $t-1$, and if $\bar{\Delta}_t$ is <u>relatively high</u> (10 or above) the <u>no stratus condition tends to continue</u>, while if $\bar{\Delta}_t$ is <u>relatively low</u> (below 10) <u>changes to a stratus condition</u> become more frequent.

These results are physically plausible, and appear consistently, if not overwhelmingly strongly, in the present data.

Figures 1 and 2 show box plots of dew point depression and ℓn(dew point depression + 1) for the years 1958-60 (cf. Tukey andMosteller (1977)). Each of the four plots in the figures contain only those points for which $y_{t-1} = i \rightarrow j = y_t$ for $i,j = 0,1$. The top (bottom) edge of the box is the upper (lower) quartile of the data set; the symbol within the box is at the median; the lines connect the mean; and the circles outside the boxes represent outlying data points.

It appears from the top two plots in each figure that dew point depression, $\bar{\Delta}_t$, may have more prognostic value if there is no stratus the day before. If there is no stratus the day before, then high $\bar{\Delta}_t$ appears to be associated with persistence of no stratus. Since the box plots do overlap, it is clear that $\bar{\Delta}_t$ will not provide perfect prediction.

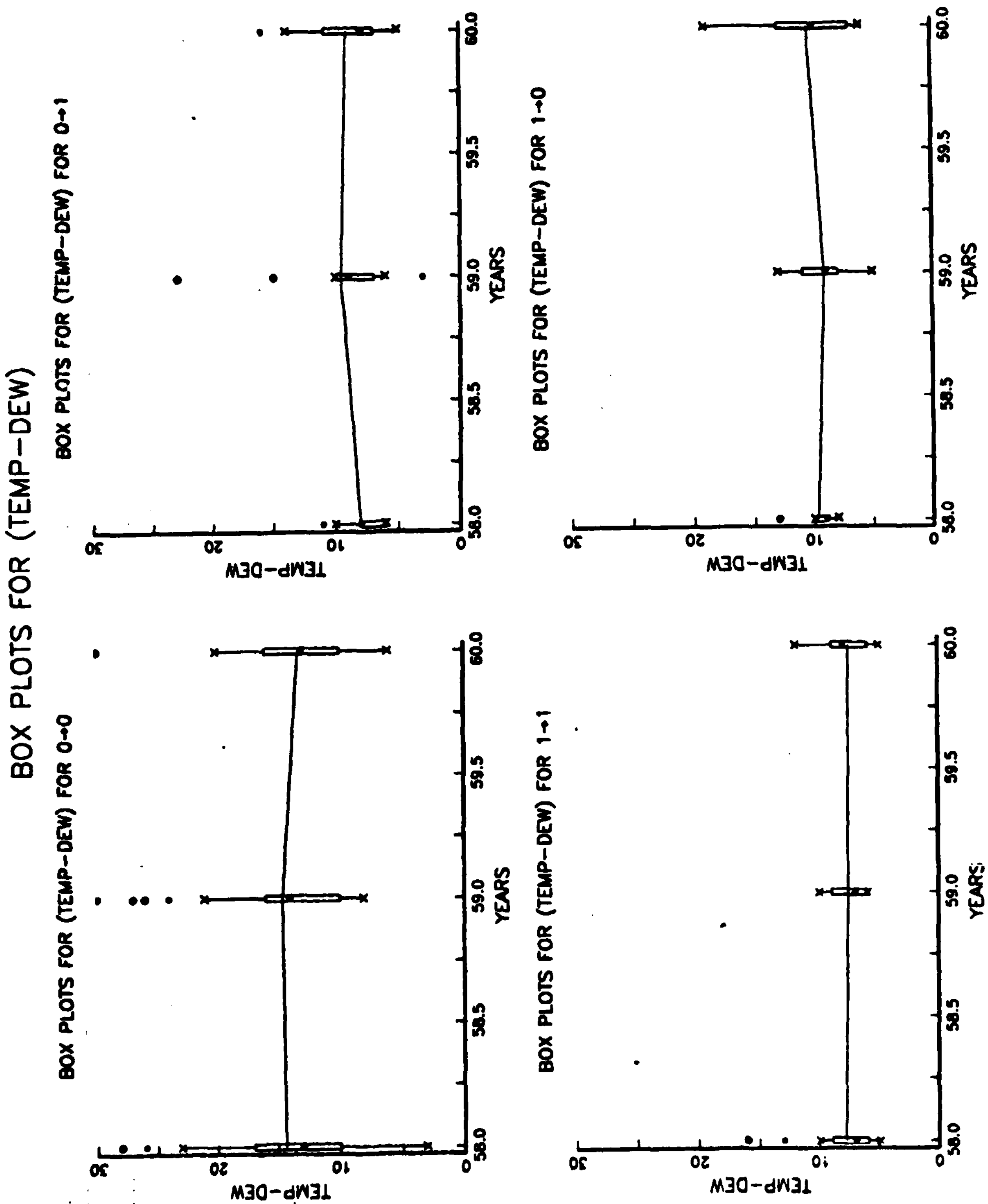

BOX PLOTS FOR (TEMP—DEW)
BOX PLOTS FOR (TEMP-DEW) FOR 0→0
BOX PLOTS FOR (TEMP-DEW) FOR 0→1
BOX PLOTS FOR (TEMP-DEW) FOR 1→1
BOX PLOTS FOR (TEMP-DEW) FOR 1→0
TEMP-DEW
YEARS
0
10
20
30
58.0
58.5
59.0
59.5
60.0

FIGURE 1.

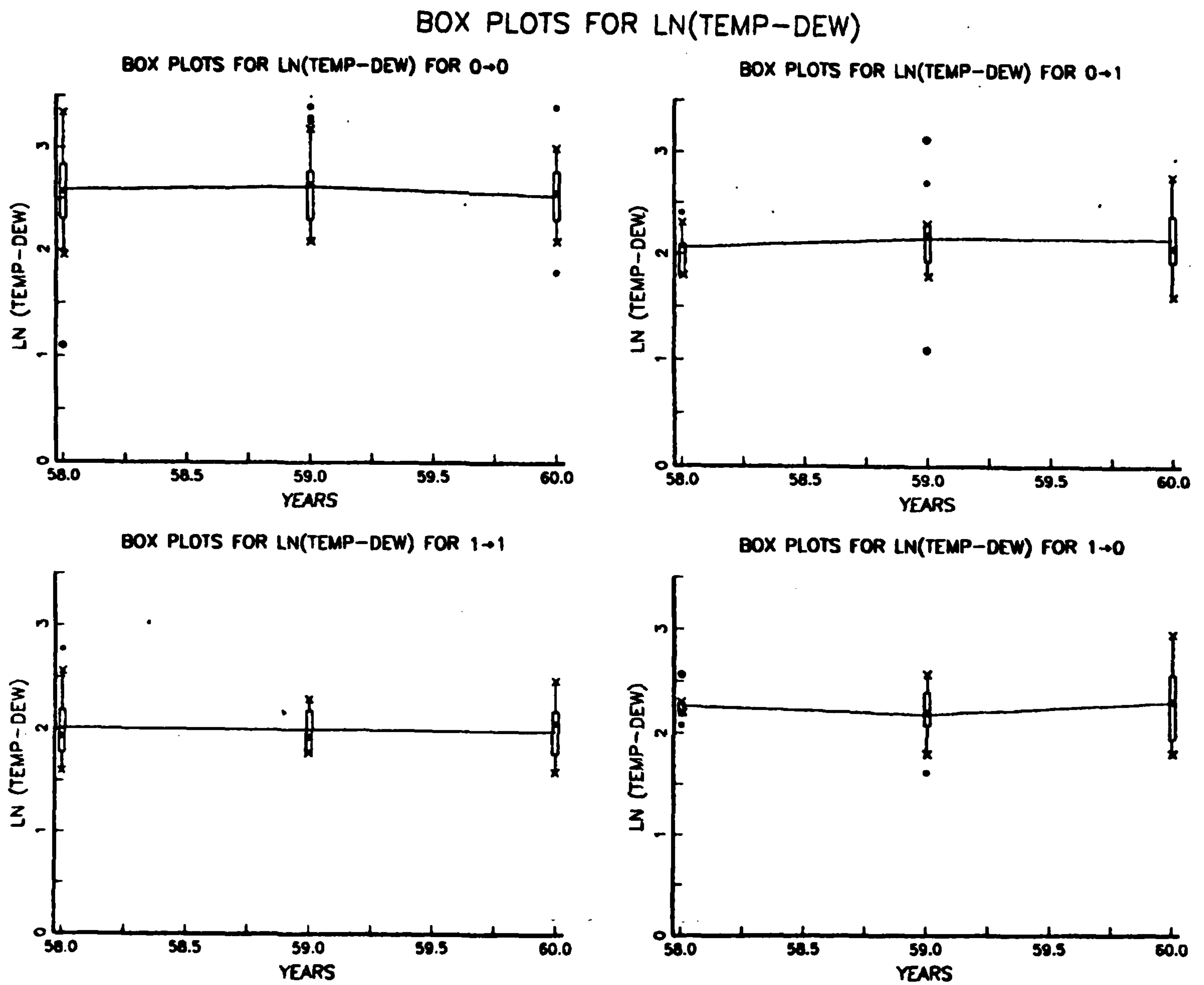

BOX PLOTS FOR LN(TEMP-DEW)
BOX PLOTS FOR LN(TEMP-DEW) FOR 0→0
BOX PLOTS FOR LN(TEMP-DEW) FOR 0→1
BOX PLOTS FOR LN(TEMP-DEW) FOR 1→1
BOX PLOTS FOR LN(TEMP-DEW) FOR 1→0
LN (TEMP-DEW)
YEARS
58.0
58.5
59.0
59.5
60.0
0
1
2
3

FIGURE 2.

An exploratory spectral analysis of hourly $\ln$(stratus height) and $\ln$(dew point depression + 1) for 1958 described in Appendix F also suggests that high (low) dew point depression is associated with high (low) stratus height.

In Table 2 are corresponding figures for hours of stratus on previous days.

TABLE 2

Observed Distribution of Previous Days' Hours of Stratus

Year		Lower Quartile ($\underline{O}$)	Median (M)	Upper Quartile ($\bar{O}$)
1958;	1 → 0:	2(0.7)	4(1.4)	8(2.1)
	1 → 1:	6(1.8)	7(1.9)	8(2.1)
1959;	1 → 0:	3(1.1)	4(1.4)	4(1.4)
	1 → 1:	4(1.4)	6(1.8)	8(2.1)
1960;	1 → 0:	3.0(1.1)	4(1.4)	4(1.4)
	1 → 1:	4(1.4)	6(1.8)	8(2.1)

Again the figures in parentheses are logs.

Again some indications from the table are of interest:

(a) corresponding summary figures are rather stable, but somewhat less so than for $\bar{\Delta}_t$,

(b) relatively low values of previous days' hours of stratus tend to be associated with change to no-stratus condition, but the tendency is rather weak.

The tendency noticed above may possibly be accounted for by the fact that an underlying weather system is passing over the Moffett area. Towards the end of its sojourn there the hours of resulting stratus tend to gradually decrease to zero.

Box plots for the number of hours of stratus the day before when there is stratus, for years 1958-60 appear in Figure 3. Each figure contains only those points for which the current day has no stratus or stratus respectively. There appears to be an association between a high number of hours of stratus the day before and persistence of stratus. The association does not appear strong, however.

Although the above sort of analysis is interesting, it fails to incorporate the joint--possibly interactive--effects of several variables. Note that no such analysis is reported here for the other possible explanatory variables related to surface wind, namely V_x and V_y. Somewhat surprisingly, these have been found to have secondary value for the location and years under investigation.

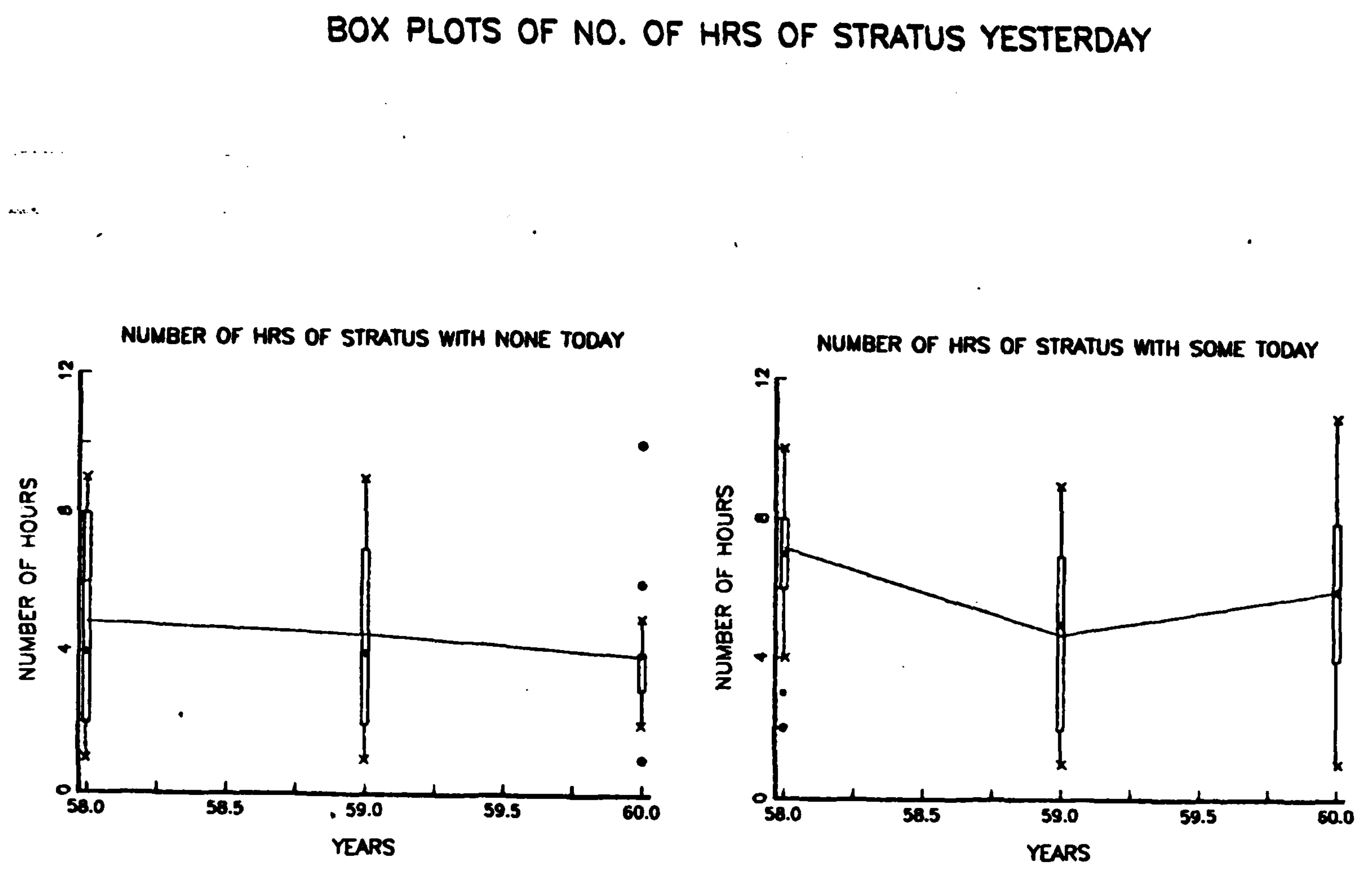

BOX PLOTS OF NO. OF HRS OF STRATUS YESTERDAY
NUMBER OF HRS OF STRATUS WITH NONE TODAY
NUMBER OF HOURS
0
4
8
12
58.0
58.5
59.0
59.5
60.0
YEARS
NUMBER OF HRS OF STRATUS WITH SOME TODAY
NUMBER OF HOURS
0
4
8
12
58.0
58.5
59.0
59.5
60.0
YEARS

FIGURE 3.

APPENDIX A

Logistic Fitting and Cross-Validation Using the BMDP Package

In Appendix E we give several mathematical justifications for use of the logistic regression model. In the present Appendix results are given of fitting various logistic models to available Moffett Field data for years 1958-60; they are cross-validated for years 1961 and 1962.

Here the term model refers to the basic logistic representation

$$P\{y=1|\text{explanatory variables } \underline{x}\} = \frac{\exp\{x\beta\}}{1+\exp\{\underline{x\beta}\}} \qquad \text{(A-1)}$$

where $\underline{x}$ is a p-vector (row) of explanatory variables, and $\underline{\beta}$ is a p-vector (column) of coefficients to be determined. The BMDP package performs the fitting, i.e. determination of $\underline{\beta}$ from observations, by maximum likelihood or a closely related method. It also furnishes Student t-values for assessing the statistical significance of the coefficients determined, and has a step-wise facility, which enters explanatory variables in accordance with their judged explanatory value. The above procedure assumes that the model is appropriate for the data, a practice that may be dangerous in observational studies, as has been pointed out by Pregibon (1983), who suggests some remedies. An examination of remedies for dealing with possibly "ill-fitting" data by the logistic is currently in progress, and will be applied to the Moffett Field, and other, data.

In the exercises reported, we have fitted 1958-1960 data by logistic models using the variable selection feature. Two types of fits are considered. In one type we condition on the

previous day's stratus state; in other words $p_0(\underline{x})$ means the probability of stratus on day t, given no stratus on $t-1$ and the influence of explanatory variables $\underline{x}$; $P_1(\underline{x})$ means the probability of stratus, given stratus on t-1. In the other type we have fit all the data at once, using an indicator variable to identify stratus - no stratus days.

The predictions made are categorical: i.e. if the calculated p-value exceeds 0.5, stratus is predicted, while if below, no stratus is predicted. We have cross-validated predictions against the years 1961 and 1962.

<u>Model A-1</u>: Prediction, given no stratus the previous day ($y_{t-1} = 0$). The explanatory variables selected are: a constant, $\ell n\,(\bar{\Delta}_t + 1)$, V_y . The fit is as follows with standard errors of the fitted parameters in parentheses below:

$$\underline{X}\hat{\beta} = \underset{(1.70)}{6.63} - \underset{(0.741)}{3.65}\,\ln(\Delta_t) - \underset{(0.0495)}{0.0878}\,V_y$$

where $\Delta_t = \bar{\Delta}_t + 1$.

The cross validations results for 1961-1962 (F means Forecast, A means Actual) are below.

<u>1961-1962</u>

A \ F	0	1	Fraction Correct
0	88	7	.93
1	16	6	.27

$$\text{Fraction Correct} = \frac{88 + 6}{88 + 6 + 7 + 16} = .80$$

Note that simple persistence forecasting ("tomorrow is the same as today") for both 1961 and 1962 gives a fraction of correct forecasts equal to 0.83 ((88+7)/(88+7+16+4)), which is actually slightly better than the logistic forecast. However, the present logistic model does correctly forecast about one-quarter to one-third of the changes from no stratus to stratus correctly; persistence will never correctly forecast a change.

Model A-2: Prediction given stratus the previous day ($y_{t-1} = 1$). The variables selected were $\ell n(\Delta_t + 1)$ and H_{t-1}. The fit is

$$\underline{x}\,\underline{\beta} = \underset{(2.07)}{6.12} - \underset{(0.940)}{3.34}\,\ell n(\Delta_t) + \underset{(0.0893)}{0.30}\,H_{t-1}$$

where $\Delta_t = \bar{\Delta}_t + 1$.

The numbers in parentheses beneath the coefficients are standard errors based upon the assumption of a correct model, maximum-likelihood fitted.

The cross validation results are below.

1961-1962

A \ F	0	1	Fraction Correct
0	16	6	.73
1	14	29	.67

$$\text{Fraction Correct} = 0.69 = \frac{16 + 29}{14 + 29 + 16 + 6}$$

In this case the logistic model did as well as persistence (0.66) in predicting stratus and no stratus. Furthermore, it predicted 73% of the changes correctly.

The results of the validation for 1961-1962 of the two fits in Exercises A-1 and A-2 are combined in the following table.

	Prediction Success	Failure	Fraction Correct
$0 \to 0$	88	7	0.93
$0 \to 1$	6	16	0.27
$1 \to 0$	16	6	0.73
$1 \to 1$	29	14	0.67

The <u>threat score</u> for predicting changes from $0 \to 1$ is

$$T_0 = \frac{C_{0\to1}}{N_{0\to1} + F_{0\to0}} = \frac{6}{(6+16+7)} = 0.21 \qquad \text{(A-2)}$$

where

$C_{0\to1}$ = the number of correct predictions of change from $0 \to 1$,

$N_{0\to1}$ = the total actual number of changes from $0 \to 1$,

$F_{0\to0}$ = the number of incorrect predictions of no change $0 \to 0$.

Similarly the threat score for predicting changes from $1 \to 0$ is

$$T_1 = \frac{C_{1\to0}}{N_{1\to0}+F_{1\to1}} = \frac{16}{16 + 6 + 14} = 0.44 \ . \qquad \text{(A-3)}$$

The threat score for predicting all changes

$$TT = \frac{C_{0\to1} + C_{1\to0}}{N_{0\to1} + N_{1\to0} + F_{0\to0} + F_{1\to1}}$$

The fraction of correct predictions using the logistic models is

$$\frac{88 + 6 + 16 + 29}{182} = 0.76.$$

The fraction of correct predictions using persistence is

$$\frac{88 + 7 + 29 + 14}{182} = 0.76.$$

Of course persistence predictions will never be correct when a change takes place, while the methods just presented, and others, may actually do quite well and seem worth the extra trouble.

Model A-3: Prediction based on all data. The variables selected are: a constant, $\ell n(\Delta_t + 1)$, and H_{t-1}. The fit is

$$\underline{x}\ \underline{\beta} = \underset{(1.30)}{6.73} - \underset{(0.570)}{3.39}\ \ell n(\Delta_t + 1) + \underset{(0.0507)}{.225}\ H_{t-1}$$

where $\Delta_t = \bar{\Delta}_t + 1$.

Again numbers in parenthesis are standard errors.

The cross-validation results are below:

1961-1962

Prediction

	Success	Failure	Fraction Correct
0 → 0	91	4	0.96
0 → 1	5	17	0.23
0 → 0	16	6	0.73
1 → 1	30	13	0.70

$$\text{Fraction Correct} = \frac{91 + 5 + 16 + 30}{182} = 0.78$$

$$\text{Fraction Correct Persistence} = \frac{91 + 4 + 30 + 13}{182} = 0.76.$$

The threat scores for predicting changes are

$$T_0 = \frac{5}{5 + 17 + 4} = 0.19$$

$$T_1 = \frac{16}{16 + 6 + 13} = 0.46$$

$$TT = \frac{16 + 5}{16 + 5 + 17 + 6 + 4 + 13} = 0.34$$

The threat scores for the fit using all the data are about the same as those for the separate fits, i.e. those that condition on whether or not stratus existed on the day before. The fraction of correct predictions of stratus and no stratus is also about the same as that for the two separate fits, and for prediction by persistence. We conclude that doing separate fits based on whether or not there is stratus the day before may not be profitable; a single logistic model may do as well as two.

APPENDIX B

Shrinkage

The term "shrinkage" is used in connection with the following phenomenon: a regression model fit by maximum likelihood (or least squares) to one set of data which is then used for prediction on another set of data nearly always fits the new set of data less well than it does the original set. Copas (1982) points out that shrinkage can be more pronounced if the original regression fit is made with the aid of a stepwise procedure; the latter tends to overfit. He suggests using the following logistic model for binary prediction:

$$P\{Y = 1 \mid \text{explanatory variable } \underline{x}\}$$

$$= \frac{\exp\{\beta_0' + K \sum_{i=1}^{n} \beta_i'(x_i - \bar{x}_i)\}}{1 + \exp\{\beta_0' + K \sum_{i=1}^{n} \beta_i'(x_i - \bar{x}_i)\}} \qquad \text{(B-1)}$$

where $\bar{x}_i$ is the mean of the i^{th} explanatory variable for the original data. $\{\beta_i'\}$ are the MLE estimators for the original data and K is a shrinkage parameter; $K = 1$ means that there is no shrinkage. Data-derived prescriptions can be found for K, but in the exploratory work reported here we have found several numerical trial values and taken note of their general effects.

The stepwise regression procedure of BMDP was used to fit a logistic model to data from 1958-60. This model with, and without shrinkage was then used to predict the occurrence of stratus in the years 1961-62. Tables 3 and 4 give the results

of the cross validation. Note that shrinkage slightly improves the prediction of no stratus on the following day.

Tables 5 and 6 give the results of fitting logistic models to the data from 1958-60 and using the models with and without shrinkage to predict stratus in 1961. Four different models were used. The parameters of the models are as follows

		Parameters
Model	A	constant, $\ell n\ \Delta_t$, y_{t-1}
	AW	constant, $\ell n\ \Delta_t$, y_{t-1}, V_x, V_y
	B	constant, $\ell n\ \Delta_t$, NS_t, NNS_t, H_{t-1}
	BW	constant, $\ell n\ \Delta_t$, NS_t, NNS_t, H_{t-1}, V_x, V_y

where Δ_t is the dew point depression plus 1.

The models were fit using maximum likelihood. Stratus was predicted on day t if the forecast probability of stratus was greater than or equal to α. The cutoff point α was taken to be 0.5, or alternatively 0.41, the fraction of days of stratus during the years 1958-1960.

Tables 7 and 8 give similar results for the models fit to data in 1958-61 and validated on 1962 data. The cutoff point α was taken to be 0.5, or alternatively 0.37, the fraction of days of stratus during the years 1958-61.

Tables 9 and 10 give the threat scores for the prediction of changes (equations A-2, A-3, and A-4).

The simplest model A with a cutoff of 0.5 seemed to do as well as any of the more complicated models. The rise of the

historical fraction of stratus days sometimes improved prediction of changes, but not in all cases. The use of shrinkage once again often seemed to improve prediction of changes from stratus to no stratus but again not uniformly. Models A and B with no shrinkage did as well as the stepwise BMDP procedure with no shrinkage.

Table 3

Validation on 61-62 of BMDP Stepwise Fit

Using all Data 58-60 with Shrinkage

K	1			0.6			0.5			0.4		
transitions	S	F	FC	S	F	FC	S	F	FC	S	F	FC
0 → 0	91	4	.96	93	2	.98	93	2	.98	95	0	1.0
0 → 1	5	17	.23	3	19	.14	3	19	.14	1	21	.05
1 → 0	16	6	.73	17	5	.77	17	5	.77	17	5	.77
1 → 1	30	13	.70	26	17	.60	26	17	.60	22	21	.51

Validation on 1961 of BMDP Stepwise Fit

Using all Data 58-60 with Shrinkage

K	1			0.6			0.5			0.4		
transitions	S	F	FC	S	F	FC	S	F	FC	S	F	FC
0 → 0	53	3	.95	54	2	.96	54	2	.96	56	0	1
0 → 1	3	8	.27	3	8	.27	3	8	.27	1	10	.09
1 → 0	8	3	.73	9	2	.82	9	2	.82	9	2	.82
1 → 1	9	4	.69	8	5	.62	8	5	.62	7	6	.54

FC = fraction correct predictions
S = number of successful predictions
F = number of unsuccessful predictions

Table 4

Validation on 1962 of BMDP Stepwise Fit

Using all Data 58-61 with Shrinkage

K	1			0.6			0.5			0.4		
transitions	S	F	FC	S	F	FC	S	F	FC	S	F	FC
0 → 0	38	1	.97	39	0	1	39	0	1	39	0	1
0 → 1	2	9	.18	0	11	0	0	11	0	0	11	0
1 → 0	8	3	.73	8	3	.73	8	3	.73	8	3	.73
1 → 1	21	9	.70	18	12	.60	16	14	.53	11	19	.37

Model has explanatory variables	constant	$\ell n(\Delta_t)$	H_{t-1}
Est coefficients	6.71	-3.42	0.242
(Std. Errors)	(1.12)	(0.489)	(0.0454)

Validation on 1962 of Separate BMDP Stepwise Fits

For Data Points with Stratus or No Stratus the Day

Before Using data of 1958-61 with Shrinkage

K	1			0.6			0.5			0.4		
transitions	S	F	FC	S	F	FC	S	F	FC	S	F	FC
0 → 0	38	1	.97	39	0	1	39	0	1	39	0	1
0 → 1	2	9	.18	0	11	0	0	11	0	0	11	0
1 → 0	8	3	.73	6	5	.55	5	6	.45	8	3	.73
1 → 1	21	9	.70	22	8	.73	24	6	.80	26	4	.86

Explanatory variables for model with no stratus the day before.

	constant	$\ell n(\Delta_t)$	V_y
Est Coefficients	6.05	-3.43	-0.081
(Std. Error)	(0.42)	(0.610)	(0.046)

Explanatory variables for model with stratus the day before.

	constant	$\ell n(\Delta_t)$	H_{t-1}
Est Coefficients	6.71	-3.56	0.291
(Std. Error)	(1.82)	(0.819)	(0.0829)

Table 5

Validation on 1961 of Predictions with Shrinkage

of Models Fit with MLE Using all Data from 1958-60

Shrinkage Parameter		Model A						Model AW					
	Cutoff	0.5			0.41			0.5			0.41		
	A\F	1	0	FC	1	0	FC	1	0	FC	1	0	FC
	1	14	10	.58	17	7	.71	11	13	.46	17	7	.71
	0	7	60	.90	13	54	.81	10	57	.85	12	55	.82
K = 1	transitions	S	F		S	F		S	F		S	F	
	0 → 0	53	3	.95	48	8	.86	50	6	.89	48	8	.86
	0 → 1	3	8	.27	5	6	.45	3	8	.27	5	6	.45
	1 → 0	7	4	.64	6	5	.54	7	4	.64	7	4	.64
	1 → 1	11	2	.85	12	1	.92	8	5	.62	12	1	.92
	A\F	1	0		1	0		1	0				
	1	11	13	.46	14	10	.58	10	14	.42	15	9	.63
	0	4	63	.94	7	60	.89	4	63	.94	11	56	.84
K = 0.6	transitions	S	F		S	F		S	F				
	0 → 0	54	2	.96	53	3	.95	54	2	.96	49	7	.88
	0 → 1	3	8	.27	3	8	.27	3	8	.27	5	6	.45
	1 → 0	9	2	.82	7	4	.64	9	2	.82	7	4	.64
	1 → 1	8	5	.62	11	2	.85	7	6	.54	10	3	.77
	A\F	1	0		1	0		1	0		1	0	
	1	8	16	.33	14	10	.58	8	16	.33	12	12	.50
	0	4	63	.94	7	60	.89	4	63	.94	11	56	.84
K = 0.5	transitions	S	F										
	0 → 0	54	2	.96	53	3	.95	54	2	.96	49	7	.88
	0 → 1	3	8	.27	3	8	.27	3	8	.27	4	7	.31
	1 → 0	9	2	.82	7	4	.64	9	2	.82	7	4	.64
	1 → 1	5	8	.38	11	2	.85	5	8	.38	8	5	.62
	A\F	1	0		1	0		1	0		1	0	
	1	5	19	.79	14	10	.58	6	18	.25	11	13	.85
	0	0	67	1	7	60	.90	1	66	.99	8	59	.88
K = 0.4	transitions	S	F		S	F		S	F		S	F	
	0 → 0	56	0	1	53	3	.93	56	0	1	52	4	.93
	0 → 1	1	10	.09	3	8	.27	2	9	.18	3	8	.27
	1 → 0	11	0	1	7	4	.64	10	1	.91	7	4	.64
	1 → 1	4	9	.31	11	2	.85	4	9	.31	8	5	.62

FC = fraction correct predictions

$$0.41 = \frac{\text{Number of Days of stratus in 1958-1960}}{\text{Total Number of Days in 1958-1960}}$$

Explanatory variables in Model A = constant, y_{t-1}, $\ln(\Delta_t)$.

Explanatory variables in Model AW = constant, y_{t-1}, $\ln(\Delta_t)$, V_x, V_y.

Table 6

Validation on 1961 of Predictions with Shrinkage

of Models Fit with MLE Using all Data from 1958-60

Shrinkage Parameter		Model B						Model BW					
	Cutoff	0.5			0.41			0.5			0.41		
K = 1	A\F	1	0	FC	1	0	FC	1	0	FC	1	0	FC
	1	13	11	.54	16	8	.67	13	11	.54	16	8	.67
	0	6	61	.91	12	55	.82	8	59	.88	12	55	.82
	transitions	S	F		S	F		S	F		S	F	
	0 → 0	53	3	.95	48	8	.86	52	4	.93	48	8	.86
	0 → 1	3	8	.27	4	7	.36	3	8	.27	5	6	.45
	1 → 0	8	3	.73	7	4	.64	7	4	.64	7	4	.64
	1 → 1	10	3	.77	12	1	.92	10	3	.77	11	2	.85
K = 0.6	A\F	1	0		1	0		1	0				
	1	11	13	.46	13	11	.54	11	13	.46	14	10	.58
	0	4	63	.94	9	58	.87	4	63	.94	11	56	.84
	transitions	S	F		S	F		S	F				
	0 → 0	54	2	.96	51	5	.91	54	2	.96	49	7	.88
	0 → 1	3	8	.27	3	8	.27	3	8	.27	3	8	.27
	1 → 0	9	2	.82	7	4	.64	9	2	.82	7	4	.64
	1 → 1	8	5	.62	12	1	.92	8	5	.62	11	2	.85
K = 0.5	A\F	1	0		1	0		1	0		1	0	
	1	11	13	.46	13	11	.54	10	14	.42	13	11	.54
	0	3	64	.96	7	60	.90	3	64	.96	11	56	.84
	transitions	S	F		S	F		S	F		S	F	
	0 → 0	55	1	.98	53	3	.95	55	1	.98	49	7	.88
	0 → 1	3	8	.27	3	8	.27	3	8	.27	3	8	.27
	1 → 0	9	2	.82	7	4	.64	9	2	.82	7	4	.64
	1 → 1	8	5	.62	10	3	.77	7	6	.54	10	3	.77
K = 0.4	A\F	1	0		1	0		1	0		1	0	
	1	7	17	.29	13	11	.54	6	18	.25	12	12	.50
	0	2	65	.97	6	61	.91	2	65	.97	8	59	.88
	transitions	S	F		S	F		S	F		S	F	
	0 → 0	56	0	1	53	3	.95	56	0	1	52	4	.93
	0 → 1	1	10	.09	3	8	.27	1	10	.09	3	8	.27
	1 → 0	9	2	.82	8	3	.73	9	2	.82	7	4	.64
	1 → 1	6	7	.46	10	3	.77	5	8	.38	9	4	.69

FC = fraction correct predictions

$$0.41 = \frac{\text{Number of Days of stratus in 1958-1960}}{\text{Total Number of Days in 1958-1960}}$$

Explanatory variables in Model B = constant, NS_t, NNS_t, H_{t-1}, $\ell n(\Delta_t)$.

Explanatory variables in Model BW = constant, NS_t, H_{t-1}, $\ell n(\Delta_t)$, V_x, V_y .

Table 7

Validation Using 1962 of Predictions Using Shrinkage

and Models Fit with MLE Using all Data from 1958-61

Shrinkage Parameter		Model A						Model AW					
	Cutoff	0.5			0.37			0.5			0.37		
K = 1	A\F	1	0	FC	1	0	FC	1	0	FC	1	0	FC
	1	26	15	.63	29	12	.71	25	16	.61	29	12	.71
	0	4	46	.93	8	42	.84	4	46	.92	7	43	.86
	transitions	S	F		S	F		S	F		S	F	
	$0 \to 0$	38	1	.97	36	3	.92	38	1	.97	37	2	.95
	$0 \to 1$	2	9	.18	3	8	.27	2	9	.18	3	8	.27
	$1 \to 0$	8	3	.73	6	5	.55	8	3	.73	6	5	.55
	$1 \to 1$	24	6	.80	26	4	.87	23	7	.77	26	4	.87
K = 0.6	A\F	1	0		1	0		1	0		1	0	
	1	15	26	.37	28	13	.68	15	26	.37	26	15	.63
	0	2	48	.96	6	44	.88	2	48	.96	5	45	.90
	transitions	S	F		S	F		S	F		S	F	
	$0 \to 0$	39	0	1	38	1	.97	39	0	1	38	1	.97
	$0 \to 1$	0	11	0	2	9	.18	0	11	0	2	9	.18
	$1 \to 0$	9	2	.82	6	5	.55	9	2	.82	7	4	.64
	$1 \to 1$	15	15	.50	26	4	.87	15	15	.50	24	6	.87
K = 0.5	A\F	1	0		1	0		1	0		1	0	
	1	15	26	.37	26	15	.63	8	33	.20	26	15	.63
	0	2	48	.96	4	46	.92	1	49	.98	5	45	.90
	transitions	S	F		S	F		S	F		S	F	
	$0 \to 0$	39	0	1	38	1	.97	39	0	1	38	1	.97
	$0 \to 1$	0	11	0	2	9	.18	0	11	0	2	9	.18
	$1 \to 0$	9	2	.82	8	3	.73	10	1	.91	7	4	.64
	$1 \to 1$	15	15	.50	24	6	.80	8	22	.27	24	6	.87
K = 0.4	A\F	1	0		1	0		1	0		1	0	
	1	7	34	.17	26	15	.63	6	35	.15	25	16	.61
	0	0	50	1.00	4	46	.92	0	50	1	4	46	.92
	transitions	S	F		S	F		S	F		S	F	
	$0 \to 0$	39	0	1	38	1	.97	39	0	1	38	1	.97
	$0 \to 1$	0	11	0	2	9	.11	0	11	0	2	9	.18
	$1 \to 0$	11	0	1	8	3	.73	11	0	1	8	3	.73
	$1 \to 1$	7	23	.23	24	6	.80	6	24	.20	23	7	.77

FC = fraction correct

$$0.37 = \frac{\text{Number of Days of stratus in 1958-1961}}{\text{Number of Days in 1958-1961}}$$

Model A explanatory variables : constant, y_{t-1}, $\ell n(\Delta_t)$.

Model AW explanatory variables: constant, y_{t-1}, $\ell n(\Delta_t)$, V_x, V_y .

Table 8

Validation Using 1962 Data of Prediction Using Shrinkage

MLE Using all Data from 1958-61 and Models Fit with

Shrinkage Parameter		Model B						Model BW					
	Cutoff	0.5			0.37			0.5			0.37		
	A\F	1	0	FC	1	0	FC	1	0	FC	1	0	FC
	1	23	18	.56	27	14	.66	23	18	.56	27	14	.66
	0	4	46	.92	8	42	.84	3	47	.94	7	43	.86
K = 1	transitions	S	F		S	F		S	F		S	F	
	0 → 0	38	1	.97	36	3	.92	39	0	1	37	2	.95
	0 → 1	2	9	.18	3	8	.27	2	9	.18	3	8	.27
	1 → 0	8	3	.73	6	5	.55	8	3	.73	6	5	.55
	1 → 1	21	9	.70	24	6	.80	21	9	.70	24	6	.80
	A\F	1	0		1	0		1	0		1	0	
	1	18	23	.44	24	17	.59	18	23	.44	24	17	.59
	0	3	47	.94	6	44	.88	3	47	.94	6	44	.88
K = 0.6	transitions	S	F		S	F		S	F		S	F	
	0 → 0	39	0	1.00	38	1	.97	39	0	1.00	38	1	.97
	0 → 1	0	11	0.00	2	9	.18	0	11	0.00	2	9	.18
	1 → 0	8	3	.73	6	5	.55	8	3	.73	6	5	.55
	1 → 1	18	12	.60	22	8	.73	18	12	.60	22	8	.73
	A\F	1	0		1	0		1	0		1	0	
	1	17	24	.41	24	17	.59	15	26	.37	24	17	.59
	0	3	47	.94	6	44	.88	3	47	.94	5	45	.90
K = 0.5	transitions	S	F		S	F		S	F		S	F	
	0 → 0	39	0	1.00	38	1	.97	39	0	1	38	1	.97
	0 → 1	0	11	0.00	2	9	.18	0	11	0	2	9	.18
	1 → 0	8	3	.73	6	5	.55	8	3	.73	7	4	.64
	1 → 1	17	13	.57	22	8	.73	15	15	.50	22	8	.73
	A\F	1	0		1	0		1	0		1	0	
	1	10	31	.24	24	17	.59	7	34	.17	23	18	.56
	0	3	47	.94	5	45	.90	3	47	.94	4	46	.92
K = 0.4	transitions	S	F		S	F		S	F		S	F	
	0 → 0	39	0	1.00	38	1	.97	39	0	1.00	38	1	.97
	0 → 1	0	11	0.00	2	9	.18	0	11	0.00	2	9	.18
	1 → 0	8	3	.73	7	4	.64	8	3	.73	8	3	.73
	1 → 1	10	20	.33	22	8	.73	7	23	.23	21	9	.70

FC = fraction correct predictions

$$0.37 = \frac{\text{Number of Days of stratus in 1958-1961}}{\text{Number of Days in 1958-1961}}$$

Model B explanatory variables: constant, NS_t, NNS_t, H_{t-1}, $\ln(\Delta_t)$.

Model BW explanatory variables: constant, NS_t, NNS_t, H_{t-1}, $\ln(\Delta_t)$, V_x, V_y

Table 9

Threat Scores for 1961 Validation
of Models fit with MLE on
data from 1958-1960

Model		A		AW		B		BW	
Cutpoint		0.5	0.41	0.5	0.41	0.5	0.41	0.5	0.41
	T_0	0.21	0.26	0.18	0.26	0.21	0.21	0.20	0.26
K = 1	T_1	0.54	0.50	0.44	0.58	0.57	0.58	0.50	0.54
	TT	0.37	0.35	0.30	0.39	0.39	0.35	0.34	0.38
	T_0	0.23	0.21	0.23	0.28	0.23	0.19	0.23	0.17
K = 0.6	T_1	0.56	0.54	0.53	0.50	0.56	0.58	0.56	0.54
	TT	0.41	0.37	0.40	0.38	0.41	0.36	0.41	0.32
	T_0	0.23	0.21	0.23	0.22	0.25	0.21	0.25	0.17
K = 0.5	T_1	0.47	0.54	0.47	0.44	0.56	0.50	0.53	0.50
	TT	0.38	0.37	0.38	0.32	0.43	0.36	0.41	0.31
	T_0	0.09	0.21	0.18	0.20	0.09	0.21	0.09	0.20
K = 0.4	T_1	0.55	0.54	0.50	0.44	0.50	0.57	0.47	0.47
	TT	0.39	0.37	0.39	0.32	0.34	0.39	0.33	0.33

		Explanatory Variables						
Model	A	constant	$\ell n(\Delta_t)$	y_{t-1}				
	AW	constant	$\ell n(\Delta_t)$	y_{t-1}	V_x	V_y		
	B	constant	$\ell n(\Delta_t)$	NS_t	NNS_t	H_{t-1}		
	BW	constant	$\ell n(\Delta_t)$	NS_t	NNS_t	H_{t-1}	V_x	V_y

$$0.41 = \frac{\text{No. days of stratus during 1958-60}}{\text{No. days during 1958-60}}$$

Table 10

Threat Scores for 1962 Validation of Models fit with MLE on data from 1958-1961

		A		AW		B		BW	
Cutpoint		0.50	0.37	0.50	0.37	0.50	0.37	0.50	0.37
K = 1	T_0	0.17	0.21	0.17	0.23	0.17	0.21	0.18	0.23
	T_1	0.47	0.40	0.44	0.40	0.40	0.35	0.40	0.32
	TT	0.34	0.31	0.33	0.32	0.31	0.29	0.32	0.30
K = 0.6	T_0	0	0.17	0	0.17	0	0.17	0	0.17
	T_1	0.35	0.40	0.35	0.41	0.35	0.32	0.35	0.32
	TT	0.24	0.30	0.24	0.31	0.24	0.26	0.24	0.26
K = 0.5	T_0	0	0.17	0	0.17	0	0.17	0	0.17
	T_1	0.35	0.47	0.30	0.41	0.33	0.32	0.31	0.37
	TT	0.24	0.34	0.23	0.31	0.23	0.26	0.22	0.29
K = 0.4	T_0	0	0.17	0	0.17	0	0.17	0	0.17
	T_1	0.32	0.47	0.31	0.44	0.26	0.37	0.24	0.40
	TT	0.24	0.34	0.24	0.33	0.19	0.29	0.18	0.31

		Explanatory Variables
Model	A	constant $\ell n(\Delta_t)$ y_{t-1}
	AW	constant $\ell n(\Delta_t)$ y_{t-1} V_x V_y
	B	constant $\ell n(\Delta_t)$ NS_t NNS_t H_{t-1}
	BW	constant $\ell n(\Delta_t)$ NS_t NNS_t H_{t-1} V_x V_y

$$0.37 = \frac{\text{No. days of stratus during 1958-61}}{\text{No. days during 1958-61}}$$

APPENDIX C

Robust Estimation for Binary Logistic Regression.

Maximum likelihood estimates are susceptible to outlying data points: they are unduly influenced by a few (exceptional) data points which may not agree with the assumed model. Pregibon (1982) suggests robust procedures which yield estimates that are resistant to a few such exceptional data points. The procedure that has been used in this report is as follows.

Let the deviance of point i be

$$d_i = -2\,(y_i \,\ell n\, \hat{p}_i + (1-y_i)\,\ell n(1-\hat{p}_i)),\ i = 1,\ldots,N \qquad \text{(C-1)}$$

where in the logistic model

$$\hat{p}_i = \frac{\exp\{\underline{x}_i\hat{\beta}\}}{1 + \exp\{\underline{x}_i\hat{\beta}\}} \qquad \text{(C-2)}$$

and

$$\underline{x}_i\underline{\hat{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}\ ; \qquad \text{(C-3)}$$

x_{ik} is the value of the $k^{\underline{th}}$ explanatory variable for the $i^{\underline{th}}$ data point, and $\hat{\beta}_k$ is the estimate of β_k, the regression coefficient for the $k^{\underline{th}}$ explanatory variable, x_k; $k = 1,\ldots,p$.

The problem of finding the MLE estimators turns out to be to solve for $\hat{\beta}_0,\ldots,\hat{\beta}_p$ in the non-linear equations

$$\sum_{i=1}^{N} x_{ik}\left[y_i - \frac{e^{\underline{x}_i\underline{\beta}}}{1 + e^{\underline{x}_i\underline{\beta}}}\right] = 0 \qquad \text{(C-4)}$$

for $k = 1,\ldots,p$.

One possible robust-resistant (insensitive to outliers) procedure is to find estimators $\hat{\beta}_0,\ldots,\hat{\beta}_p$ such that

$$\sum_{i=1}^{N} w(i)x_{ik}\left[y_i - \frac{e^{\underline{x}_i\underline{\beta}}}{i + e^{\underline{x}_i\underline{\beta}}}\right] = 0 \qquad \text{(C-5)}$$

where

$$w(i) = \begin{cases} 1 & \text{if} \quad d_i \leq H \\ (H/d_i)^{1/2} & \text{otherwise,} \end{cases} \qquad \text{(C-6)}$$

d_i is the deviance of the $i^{\underline{th}}$ data point and the fitted model at that point, from (C-1).

A value of $H = 1.35$ was suggested by Pregibon and used for the tuning constant; if $H = \infty$ the procedure carries out the ordinary MLE fitting, while as H decreases the effects of extreme local deviance points have progressively less effect on the fitted model. Notice that the $i^{\underline{th}}$ data-determined weight, $w(i)$, is made relatively small if $d(i)$ is large. Thus data points which are not well fit by the assumed model will tend to receive less weight than others that are. The resistant estimates, $\hat{\underline{\beta}}$, are found by iteration. First the MLE estimate is found and the initial weights computed. Then (C-5) is solved for $\{\hat{\beta}_k(1), k = 1,\ldots,p\}$ by a Newton-Raphson procedure. New weights $w_k(1)$ are computed from (C-6). Then these are entered in (C-5), and it is solved for $\{\hat{\beta}_k(2), k = 1,\ldots,p\}$; this process repeats until the iterative estimates converge.

On each day either stratus occurs or not. If stratus occurs on consecutive days then a run of stratus days is said to occur. Let NS_t be the length of the run of stratus days that includes day t-1. For example, $NS_t = 0$ if the previous day had no stratus, so $y_{t-1} = 0$; while $NS_t = 2$ if $y_{t-1} = 1$, $y_{t-2} = 1$, $y_{t-3} = 0$. Let NNS_t be the length of the last run of no stratus days that includes day $t - 1$.

Table (11) gives the estimates for five iterations of the robust procedure applied to a model using 1958-1960 data. The explanatory variables are: constant, NS_t, NNS_t, H_{t-1}, $\ell n(\Delta_t)$. where Δ_t is the dew point depression plus 1.

TABLE 11

Results of Iteration of Resistant Procedure

Number of Iteration	Constant	NS_t	NNS_t	H_{t-1}	$\ell n(\Delta_t)$
0 (MLE)	6.81	-0.01	-0.05	0.21	-3.34
1	9.30	-0.04	-0.05	0.28	-4.55
2	9.98	-0.05	-0.04	0.30	-4.88
3	10.16	-0.06	-0.04	0.31	-4.97
4	10.21	-0.06	-0.04	0.31	-5.00
5	10.22	-0.06	-0.04	0.31	-5.00

Note that except for the estimated value of NNS_t the resistant procedure has made the estimates greater in absolute value. Such sharpening of the expression is a common occurrence when robust logistic procedures are utilized.

We fit this model B robustly to 1958-60 data and then used the fitted model to predict the occurrence of stratus with a cutoff point of 0.5. We also robustly fit model B to 1958-61 data and used it to predict the occurrence of stratus in 1962. Although the estimated parameters using the robust procedure were different, the results of the cross-validation were almost the same as with the maximum likelihood fit reported in Appendix B. Results of the cross-validations with models fit robustly appear in Table 18 at the end of Appendix D.

APPENDIX D

Logistic Models with Updating

Despite best attempts to develop a single model with which to predict stratus in any given year, the resulting model may suffer from lack of timeliness. The basic reason is that simple models fitted with data from one period may well not be entirely relevant to another, owing to changing conditions not represented in the model. One attractive procedure for dealing with the lack of timeliness issue is to progressively update the model fit so as to incorporate recent data, i.e. data representing conditions near in time to those to be forecast. This is the philosophy of the well-known Kalman filter. In the present context the updating procedure has been carried out completely straightforwardly, i.e. by simply re-computing estimates using recent data. Computationally economical and sophisticated methods remain to be developed.

We report the results of an investigation of updated model fits to predict the occurrence of stratus. Three updating schemes were tried.

I. A model was initially fit using all data from the previous year. Then a forecast of the occurrence of stratus was made using the model for the first ten days of the current (forecast) year. These ten days were then added to the forecasting data set, and the eldest, or initial, ten days of data were dropped. The model was re-fit using the updated data. Using the new model, the occurrence of stratus the next ten days of the current year was forecast. Then the second-eldest ten-days-worth of data were dropped, and the newest ten days were added, and the model was

re-fit, forecasts made, and so the process was continued. This may be referred to as a 90-day rolling forecast in steps of ten days.

II. A model was initially fit using all data from the previous year. A forecast for the occurrence of stratus was made for the first day of the current year. This data point was added to the forecasting data set, and the eldest point deleted. The model was refit using the altered modeling data set. A forecast of the occurrence of stratus was made for the next day of the current year. This data point was added to the modeling data set and the oldest point was dropped, and so forth. This is a rolling forecast in one-day steps.

III. Same as II but the initial modeling data set includes only the last 45 points of the previous year.

Two different sets of explanatory variables were tried, A and B with and without wind speeds, where

A: constant, y_{t-1}, $\ell n(\Delta_t)$.

AW: constant, y_{t-1}, $\ell n(\Delta_t)$, $V_x(t)$, $V_y(t)$

B: constant, NS, NNS_t, H_{t-1}, $\ell n(\Delta_t)$

BW: constant, NS_t, NNS_t, H_{t-1}, $\ell n(\Delta_t)$, $V_x(t)$, $V_y(t)$

as before; Δ_t is the dew point depression plus 1.

A prediction of stratus was made if the forecasted probability was greater than α. In most cases $\alpha = 0.5$. Additionally, α was sometimes taken to be the fraction of the number of days of stratus over all years previous to the current year.

The results are summarized in Tables 12-14 of threat scores (T_0, T_1, TT) and fraction of correct predictions (FC). For comparison purposes results are also given for prediction without

updating. Full tables of the numbers of correct and incorrect predictions can be found in Tables 15-18.

As stated previously, the cutoff point, α , for the updating procedures was either 0.5, or alternatively, the historical fraction of days of stratus. For the simpler model A, the use of the historical fraction appeared to improve prediction of stratus, but to worsen the prediction of no stratus. Using robust estimates in updating procedure I gave the same results as using the simpler MLE estimates. The more complicated model B often (but not always) improved predictions of changes. Adding information about winds to either model A or B never improved prediction much. Using shrinkage with the updating procedure II once again tended to improve prediction of changes from stratus to no stratus, but tended to worsen prediction of a change from no stratus to stratus. Updating procedure III often seemed to do better in predicting changes from no stratus to stratus than updating procedure II; however, it did worse when predicting changes from stratus to no stratus. Updating procedure I always did at least as well as in predicting changes from stratus to no stratus but sometimes not as well as III in predicting changes from no stratus to stratus. Model B with an updating procedure often did better than Model A with updating particularly in predicting changes from no stratus to stratus. In summary, models with updating sometimes did better than models with no updating, but the improvement was surprisingly small.

Table 12

Threat Scores for Changes and Fraction of Predictions Correct for 1961 Predictions

Based on Models With and Without Updating

Model A

Data Used to Fit Model	1958-1960		1960					AW 1958-1960	
Updating	NO		NO	I		II	III	NO	
Method	MLE		MLE	MLE		MLE	MLE	MLE	
α	0.5	0.41	0.5	0.5	0.41	0.5	0.5	0.5	0.41
T_0	0.21	0.26	0.26	0.25	0.39	0.25	0.29	0.18	0.26
T_1	0.54	0.50	0.56	0.53	0.50	0.50	0.50	0.44	0.58
TT	0.37	0.35	0.40	0.39	0.44	0.38	0.39	0.30	0.39
FC	0.81	0.78	0.77	0.79	0.80	0.78	0.78	0.75	0.79

Model B

Data Used to Fit Model	1958-1960			1960					BW	
Updating	NO			NO	I		II	III	1958-1960	
Method	MLE		Robust	MLE	MLE	Robust	MLE	MLE	MLE	
α	0.5	0.41	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.41
T_0	0.21	0.21	0.21	0.35	0.43	0.43	0.29	0.29	0.20	0.26
T_1	0.57	0.58	0.57	0.56	0.60	0.60	0.56	0.44	0.50	0.54
TT	0.39	0.35	0.39	0.45	0.52	0.52	0.43	0.36	0.34	0.38
FC	0.81	0.78	0.81	0.80	0.84	0.84	0.81	0.77	0.79	0.78

Model A has explanatory variables: constant, y_{t-1}, $\ell n(\Delta_t)$

Model B has explanatory variables: constant, NS_t, NNS_t, H_{t-1}, $\ell n(\Delta_t)$

Fraction correct using persistence is 0.76

Table 13

Threat Scores for changes and fraction of predictions correct for 1962 Predictions

Model A

Data Used to Fit Model	1958-1961		1961					AW 1958-1961	
Updating	NO		NO	I		II	III	NO	
Method	MLE		MLE	MLE		MLE	MLE	MLE	
α	0.5	0.37	0.5	0.5	0.37	0.5	0.5	0.5	0.37
T_0	0.17	0.21	0	0	0.21	0	0.14	0.17	0.23
T_1	0.47	0.40	0.47	0.31	0.17	0.31	0.27	0.44	0.40
TT	0.34	0.31	0.29	0.15	0.19	0.15	0.21	0.33	0.32
FC	0.79	0.78	0.78	0.74	0.77	0.76	0.75	0.78	0.79

Model B

Data Used to Fit Model	1958-1961			1961					BW	
Updating	NO			NO	I		II	III	1958-1961	
Method	MLE		Robust	MLE	MLE	Robust	MLE	MLE	MLE	
α	0.5	0.37	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.37
T_0	0.17	0.21	0.17	0.17	0.15	0.15	0.08	0.20	0.18	0.23
T_1	0.40	0.40	0.35	0.29	0.37	0.37	0.26	0.22	0.40	0.35
TT	0.31	0.31	0.28	0.24	0.28	0.28	0.19	0.21	0.32	0.30
FC	0.76	0.76	0.75	0.73	0.74	0.74	0.71	0.71	0.77	0.77

Model A has explanatory variables: constant, y_{t-1}, $\ell n\ (\Delta_t)$

Model B has explanatory variables: constant, NS_t, NNS_t, H_{t-1}, $\ell n\ (\Delta_t)$

Fraction correct using persistence is 0.76

Table 14

Threat Scores for Validations on 1961

K	B Updating 1 day at a time With Shrinkage Start with all of 1960 Cutoff = 0.5: 1	0.6	0.5	A 10 da Rolling Cutoff = 0.5	(Results for B)	A 10 da Rolling Cutoff = 0.41	A Update 1 day at a time Cutoff = 0.5 E (90 days)	H (45 days)	B Update 1 day at a time Cutoff = 0.5 E	H	A Fit On 60 Cutoff = 0.5	A Fit On 58-60 Cutoff = 0.5	B Fit On 60 Cutoff = 0.5	B Fit On 58-60 Cutoff = 0.5	BMDP One Fit On 58-60 Cutoff = 0.5
T_0	0.29	0.23	0.17	0.25	(0.43)	0.39	0.25	0.29	0.29	0.29	0.26	0.21	0.35	0.21	.21
T_1	0.56	0.53	0.56	0.53	(0.60)	0.50	0.50	0.50	0.56	0.44	0.56	0.54	0.56	0.57	.53
TT	0.43	0.40	0.40	0.39	(0.52)	0.44	0.38	0.39	0.43	0.36	0.40	0.37	0.45	0.39	.38
FC	0.81	0.80	0.80	0.79	(0.84)	0.80	0.78	0.78	0.81	0.77	0.77	0.81	0.80	0.81	0.80

Threat Scores for Validations on 1962

K	B Updating 1 day at a time With Shrinkage Start with all of 1961 Cutoff = 0.5: 1	0.6	0.5	A 10 da Rolling Cutoff = 0.5	(Results for B)	A 10 da Rolling Cutoff = 0.37	A Update 1 day at a time Cutoff = 0.5 E	H	B Update 1 day at a time Cutoff = 0.5 E	H	A Fit On 61 Cutoff = 0.5	A Fit On 58-61 Cutoff = 0.5	B Fit On 61 Cutoff = 0.5	B Fit On 58-61 Cutoff = 0.5	BMDP One Fit On 58-61 Cutoff = 0.5
T_0	0.08	0	0	0	(0.15)	0.21	0	0.14	0.08	0.20	0	0.17	0.17	0.17	.17
T_1	0.26	0.32	0.36	0.31	(0.37)	0.17	0.31	0.27	0.26	0.22	0.47	0.47	0.29	0.40	.40
TT	0.19	0.21	0.24	0.15	(0.28)	0.19	0.15	0.21	0.19	0.21	0.29	0.34	0.24	0.31	.31
FC	0.76	0.70	0.71	0.74	(0.74)	0.77	0.76	0.75	0.71	071	0.78	0.79	0.73	0.76	0.76

E = update starts with a model fit with entire previous year.

H = update starts with a model fit with half of previous year (last 45 days).

Model		explanatory variables
	A	constant, $\ell n(\Delta_t)$, y_{t-1}
	B	constant, $\ell n(\Delta_t)$, NS_t, NNS_t, H_{t-1} .

The fraction of correct predictions of daily stratus or no stratus made during both years using only persistence is 0.76.

Table 15

Validations For Rolling Fits
(10 days at a time initiating with only the previous year)

	Model (Fitting Procedure)		A (MLE)						B (MLE)			B (Robust)		
	Cutoff		0.5			α			0.5			0.5		
1958-59			0.5			0.52								
		A\F	1	0	FC	1	0	FC	1	0	FC	1	0	FC
	Stratus	1	12	12	.50	11	13	.46	14	10	.58	14	10	.58
	No Stratus	0	8	58	88	8	58	.88	8	58	.88	8	58	.88
			S	F		S	F		S			S	F	
		0 → 0	51	0	1	51	0	1	50	1	.98	50	1	.98
	transitions	0 → 1	2	12	.17	2	12	.17	5	9	.36	5	9	.36
		1 → 0	7	8	.47	7	8	.47	8	7	.53	8	7	.53
		1 → 1	10	0	1	9	1	.90	9	1	.90	9	1	.90
1959-60			0.5			0.267			0.5			0.5		
		A\F	1	0		1	0		1	0		1	0	
		1	22	19	.54	36	5	.88	20	21	.49	20	21	.49
		0	9	40	.82	22	27	.55	6	43	.88	6	43	.88
			S	F		S	F		S	F		S	F	
		0 → 0	31	3	.91	23	11	.68	31	3	.91	31	3	.91
		0 → 1	5	10	.67	10	5	.67	4	11	.27	4	11	.27
		1 → 0	9	6	.60	4	11	.27	12	3	.80	12	3	.80
		1 → 1	17	9	.65	26	0	1	16	10	.62	16	10	.62
1960-61			0.5			0.41			0.5			0.5		
		A\F	1	0		1	0		1	0		1	0	
		1	13	11	.54	17	7	.71	15	9	.63	15	9	.63
		0	8	58	.88	11	55	.83	5	61	.92	5	61	.92
			S	F		S	F		S	F		S	F	
		0 → 0	50	5	.91	48	7	.87	52	3	.95	52	3	.95
		0 → 1	4	7	.36	7	4	.64	6	5	.55	6	5	.55
		1 → 0	8	3	.73	7	4	.64	9	2	.82	9	2	.82
		1 → 1	9	4	.69	10	3	.77	9	4	.69	9	4	.69
1961-62			0.5			0.37			0.5			0.5		
		A\F	1	0		1	0		1	0		1	0	
		1	28	13	.68	32	9	.78	24	17	.59	24	17	.59
		0	10	39	.80	12	37	.76	6	43	.88	6	43	.88
			S	F		S	F		S	F		S	F	
		0 → 0	35	3	.92	35	3	.92	36	2	.95	36	2	.95
		0 → 1	0	11	0	3	8	.27	2	9	.18	2	9	.18
		1 → 0	4	7	.36	2	9	.18	7	4	.64	7	4	.64
		1 → 1	28	2	.93	29	1	.97	22	8	.73	22	8	.73

$$\alpha = \frac{\text{number of days of stratus during all previous years}}{\text{number of days in all previous years}}$$

Model A explanatory variables: constant, y_{t-1}, $\ell n(\Delta_t)$

Model B explanatory variables: constant, NS_t, NNS_t, H_{t-1}, $\ell n(\Delta_t)$

Table 16

One year Validations for Updating MLE fits
of models for one day ahead and dropping
the oldest day (cutoff = 0.5)

	Model	A						B					
	Initiating Data Set	E			H			E			H		
	A\F	1	0	FC	1	0	FC	1	0	FC	1	0	FC
	1	11	13	.46	8	16	.33	15	9	.63	8	16	.33
1958-59	0	8	59	88	5	62	.93	8	59	.88	10	57	.85
	transitions	S	F		S	F		S	F		S	F	
	$0 \to 0$	52	0	1	50	2	.96	51	1	.98	47	5	.76
	$0 \to 1$	2	12	.14	2	12	.14	5	9	.64	4	10	.94
	$1 \to 0$	7	8	.47	12	3	.80	8	7	.53	10	5	.67
	$1 \to 1$	9	1	.90	6	4	.60	10	0	1	4	6	.40
	A\F	1	0		1	0		1	0		1	0	
	1	24	17	.59	24	17	.59	22	19	.54	24	17	.59
1959-60	0	10	40	.80	12	38	.76	8	42	.84	11	39	.78
		S	F		S	F		S	F		S	F	
	$0 \to 0$	32	2	.94	31	3	.91	32	2	.94	31	3	.91
	$0 \to 1$	5	10	.33	4	11	.27	4	11	.27	4	11	.27
	$1 \to 0$	8	8	.50	7	9	.44	10	6	.63	8	8	.50
	$1 \to 1$	19	7	.73	20	6	.77	18	8	.69	20	6	.77
	A\F	1	0		1	0		1	0		1	0	
	1	12	12	.50	13	11	.54	12	12	.50	13	11	.54
1960-61	0	8	59	.88	9	58	.87	5	62	.93	10	57	.85
		S	F		S	F		S	F		S	F	
	$0 \to 0$	51	5	.91	50	6	.89	53	3	.95	50	6	.89
	$0 \to 1$	4	7	.36	5	6	.45	4	7	.36	5	6	.46
	$1 \to 0$	8	3	.73	8	3	.73	9	2	.82	7	4	.64
	$1 \to 1$	8	5	.62	8	5	.62	8	5	.62	8	5	.62
	A\F	1	0		1	0		1	0		1	0	
	1	28	13	.68	28	13	.68	23	18	.56	26	15	.63
1961-62	0	9	41	.82	10	40	.80	8	42	.84	11	39	.78
		S	F		S	F		S	F		S	F	
	$0 \to 0$	37	2	.95	36	3	.92	37	2	.95	35	4	.90
	$0 \to 1$	0	11	0	2	9	.18	1	10	.09	3	8	.27
	$1 \to 0$	4	7	.36	4	7	.36	5	6	.45	4	7	.36
	$1 \to 1$	28	2	.93	26	4	.87	22	8	.73	23	7	.77

E: entire previous year used to fit initial model

H: half previous year used to fit initial model

Model A explanatory variables: constant, y_{t-1}, $\ell n(\Delta_t)$

Model B explanatory variables: constant, NS_t, NNS_t, H_{t-1}, $\ell n(\Delta_t)$

FC = fraction correct predictions

Table 17

Validation of Updating of Model B with Shrinkage. The Model was initially fit with entire previous year and one point from new year added and oldest point dropped in each iteration.

	K	1			0.6			0.5			0.4		
	A\F	1	0	FC	1	0	FC	1	0	FC	1	0	FC
1960-61	1	12	12	.50	10	14	.42	8	16	.33	5	19	.21
	0	5	62	.93	4	63	.94	2	65	.97	1	66	.99
	transitions	S	F		S	F		S	F		S	F	
	0 → 0	53	3	.95	54	2	.96	55	1	.98	56	0	1
	0 → 1	4	7	.36	3	8	.27	2	9	.18	1	10	.09
	1 → 0	9	2	.82	9	2	.82	10	1	.91	10	1	.91
	1 → 1	8	5	.62	7	6	.54	6	7	.46	4	9	.31
	A\F	1	0		1	0		1	0		1	0	
1961-62	1	23	18	.56	19	22	.46	19	22	.46	15	26	.37
	0	8	42	.84	5	45	.90	4	46	.92	4	46	.92
	transitions	S	F		S	F		S	F		S	F	
	0 → 0	37	2	.95	38	1	.97	38	1	.97	38	1	.97
	0 → 1	1	10	.09	0	11	0	0	11	0	0	11	0
	1 → 0	5	6	.45	7	4	.64	8	3	.73	8	3	.73
	1 → 1	22	8	.73	19	11	.63	19	11	.63	15	15	.50

FC = fraction correct

Model B explanatory variables: constant, NS_t, NNS_t, H_{t-1}, $\ln(\Delta_t)$.

Cutoff = 0.5 .

Table 18

Validations for MLE fits without updating based on different amounts of historical data

	Validation Yr.	1961						1962					
	Historical data	1960			1958-60			1961			1958-61		
	A\F	1	0	FC	1	0	FC	1	0	FC	1	0	FC
Model A	1	13	11	.54	14	10	.58	24	17	.59	26	15	.63
	0	10	57	.85	7	60	.90	3	47	.94	4	46	.92
	transitions	S	F		S	F		S	F		S	F	
	0 → 0	48	8	.86	53	3	.95	39	0	1	38	1	.97
	0 → 1	5	6	.45	3	8	.27	0	11	0	2	9	.18
	1 → 0	9	2	.82	7	4	.64	8	3	.73	8	3	.73
	1 → 1	8	5	62	11	2	85	24	6	.80	24	6	.80
	A\F	1	0	FC	1	0	FC	1	0	FC	1	0	FC
Model B	1	14	10	.58	13	11	.54	22	19	.54	23	18	.56
	0	8	59	.88	6	61	.91	6	44	.88	*4	46	.92
	transitions	S	F		S	F		S	F		S	F	
	0 → 0	50	6	.89	53	3	.95	38	1	.97	38	1	.97
	0 → 1	6	5	.55	3	8	.27	2	9	.18	2	9	.18
	1 → 0	9	2	.82	8	3	.73	6	5	.55	+8	3	.73
	1 → 1	8	5	62	10	3	.77	20	10	.67	21	9	.70

Cutoff point = 0.5 .

Model B fit robustly to data 1958-60 and cross-validated on 1961 gives the same results as MLE.

Model B fit robustly to data 1958-61 and cross-validated on 1962 gives the same results as MLE except in the cases * and +; for * the corresponding numbers are 5 and 45; for + the corresponding numbers are 7 and 4.

APPENDIX E

Survival Models: Relation to the Logistic Representation.

E.1 Preliminary Models

Suppose a system occupies one of two states for a varying ("random") time period, then switches to the other, and back. Such events occur at times $t = 0,1,2,3,\ldots$. Such is the case with the stratus-no stratus fluctuation that has been studied, but is also true of many other weather-related events, rainfall-no rainfall being a prime example.

We discuss several traditional stochastic models as a preliminary.

Model 1: Markov Chain

Let Y_t denote the state variable of the system at time t. Suppose (here $i,j = 0,1$)

$$P\{Y_t=j \mid Y_{t-1}=i\} = p_{ij} > 0 \; ; \tag{E-1}$$

in particular, no further past history is useful:

$$P\{Y_t=j \mid Y_{t-1}=i, Y_{t-2}=a, Y_{t-3}=b, \ldots, Y_{t-\ell}=k \ldots\} = p_{ij} \tag{E-2}$$

for all i,j and all t.

There is then a long-run or steady-state distribution $\{\pi_0, \pi_1\}$ that satisfies balance equations:

$$\pi_0 p_{01} = \pi_1 p_{10} = (1-\pi_0) p_{10} \tag{E-3}$$

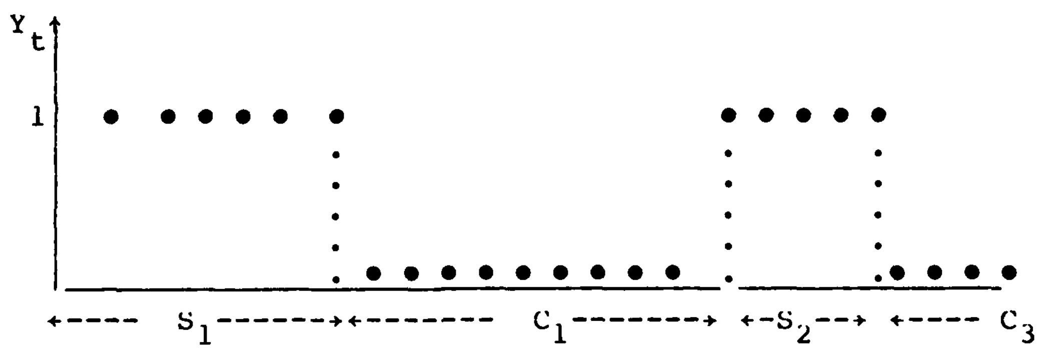
so

$$\pi_0 = \frac{p_{10}}{p_{10} + p_{01}}, \quad \pi_1 = \frac{p_{01}}{p_{10} + p_{01}}$$

If such a model truly described nature, i.e. stratus level at an airport, then π_1 could be referred to as the climatological probability of stratus, $(Y_t=1)$, on a day. Such a model may be fitted to data: one simply estimates p_{10}, for example, by the fraction of changes from 1 to 0 (stratus to no-stratus) observed in an observational period. The model does not have the capacity to incorporate physical parameters or explanatory variables, such as dewpoint depression.

Model 2: Two-State Renewal Process

Let S represent the generic length of a stratus period, i.e. or number of days throughout which there is uninterrupted stratus $(Y_t=1)$. Just before S, and just after, there will be periods of one or more no-stratus days; let such a generic period be C (C denotes "clear"); (throughout the period $Y_t = 0$). If $\{S_i\}$ is a sequence of statistically indpendent stratus periods from the same distribution, and $\{C_i\}$ is a collection of corresponding clear periods, then the time history of system state appears as below:



In the long run,

$$\lim_{t \to \infty} P\{Y_t = 1\} = \frac{E[S]}{E[S] + E[C]}$$

$$= \frac{\text{Mean Length of Stratus Period}}{\text{Mean Length of Strat. + Mean Length of No-Strat.}} \qquad \text{(E-5)}$$

The above can be called the climatological probability of stratus on a day. Strictly, the two-state renewal process model stipulates that the sequence of stratus day periods $\{S_j\}$ is one of independent, identically distributed random variables, as is the sequence of clear day periods $\{C_k\}$; the two sequences are mutually independent. The Markov chain model is a special case of the two-state renewal model in which stratus periods, generically S, have a geometric distribution with mean $E[S]$, and the clear periods, C, have their, generally different, geometric distribution with mean $E[C]$.

Once again, this model contains no direct accounting for the possible influence of explanatory variables upon the probabilities of stratus state changes.

E.2 The General Survival Model

Suppose a forecaster is in action at time t. He easily notes the current system state; suppose $Y_t = 0$, i.e. no stratus. He wishes to predict the system state at $t + 1$. A believer in Model 2 will act in an actuarial fashion, computing the conditional probability that the same state will prevail ("survival" occurs), given that the current clear state has lasted for d days:

$$P\{C \geq d+1 | C \geq d\} = e^{-h_0(d+1)} \quad \text{(E-6)}$$

or

$$P\{Y_{t+1}=1 | Y_t=0, Y_{t-1}=0, \ldots, Y_{t-d+1}=0, Y_{t-d}=1, \ldots\} =$$

$$= P\{C<d+1 | C \geq d\} \quad \text{(E-7)}$$

$$= 1 - e^{-h_0(d+1)} .$$

Similarly, if stratus is now present $(Y_t=1)$,

$$P\{Y_{t+1}=1 | Y_t=1, Y_{t-1}=1, \ldots, Y_{t-d+1}=1, Y_{t-d}=0, \ldots\} \quad \text{(E-8)}$$

$$= e^{-h_1(d+1)} ;$$

the quantities $h_0(d)$, $h_1(d)$ may be referred to as the <u>hazards</u> associated with the states in question, for

$$1 - e^{-h_1(d+1)} \simeq h_1(d+1) \quad \text{if} \quad h_1(d+1) \quad \text{is small}$$

is the conditional probability, or, picturesquely, <u>hazard</u>, that a stratus period of duration ("lifelength" or "age") d actually "dies", or changes to a non-stratus period at age $d + 1$. Similarly when a non-stratus period is in progress, the change occurs with hazard $h_0(d+1)$.

A promising enterprise is now to enhance the above forecast of survival, or death, at age $d + 1$ by further relevant information about the physical environment of the process. Under present circumstances, i.e. when forecasting stratus, one might well use

dewpoint depression Δ_t as well as previous days of stratus (or no-stratus). Other explanatory variables might well be appropriate, and can perhaps be identified from physical arguments augmented by graphical or other exploratory techniques.

In order to utilize the hazard notion in a regression context it is convenient to put

$$h_0 = \exp\{\underline{x}_t \underline{\beta}_0\} \tag{E-10}$$

where for instance the vector of explanatory variables might be

$$\underline{x}_t = (1, NS_t, \Delta_t, H_{t-1}, t) \tag{E-11}$$

and

$$\underline{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p}) \tag{E-12}$$

is the required system of constants. A form such as (E-10) can never be negative, a minimal requirement. Precisely the expression (E-10) has been used by Cox (1972) for describing hazards. Actually Cox's hazard is written as

$$\lambda(t)\exp\{\underline{x\beta}\} \tag{E-13}$$

Suppose observations are available on n days: these are of the form

$$(y_t, x_{t1}, x_{t2}, \ldots, x_{tp}) \; ,$$

where, as was mentioned earlier, possibly

$x_{t1} = 1$, $x_{t2} = \ln(\Delta_t)$, $x_{t3} = H_{t-1}$, $x_{t4} = NS_t$(i.e. # days of continuous stratus) .

Note that interactions and transformations can directly be included; e.g. simply put $x_{t5} = x_{t3}x_{t2} = H_{t-1} \times \ln(\Delta_t)$ to represent an interaction term.

Now the likelihood for the $\underline{\beta}_0$ vector is

$$L(\underline{\beta}_0;\underline{y},\underline{x}) = \prod_{t=1}^{n} [e^{-h_0(\underline{x}_t)}]^{y_t}[1-e^{-h_0(\underline{x}_t)}]^{1-y_t}; \tag{E-14}$$

taking logs, we get

$$\ell(\underline{\beta}) = \sum_{t=1}^{n} [y_t h_0(\underline{x}_t) + (1-y_t)\ln[1-e^{-h_0(\underline{x}_t)}]]$$

$$= - \sum_{t=1}^{n} [y_t \exp[\underline{x}_t\underline{\beta}_0] + (1-y_t)\ln(1-\exp\{\underline{x}_t\underline{\beta}_0\})] \tag{E-15}$$

and this can be maximized by choice of $\underline{\beta}_0$, a non-linear optimization task. The usual approach would involve differentiation with respect to β_{0j}, and solving the resulting non-linear system by a variation of the Newton-Raphson method. Package programs are available for such a task.

E.3 The Logistic Model from Cox's Model

Suppose a Cox model is under consideration for describing the probability distribution of "age to death" or, in the present context, the survival of a stratus (or no-stratus) episode for another day. In a simple form, the probability of survival through $t + 1$ in state j ($j = 1,0$) given that for the past m time periods state j is in effect is

$$P_j(m,\underline{x}_t) \equiv P\{Y_{t+1}=j \mid Y_t=j, Y_{t-1}=j, \ldots \; Y_{t-(m-1)}=j, Y_{t-m}\neq j, \underline{X}_t = \underline{x}_t\}$$

$$= e^{-h_j(m+1)} = \exp[-\lambda_j(t)\exp\{\underline{x}_t\underline{\beta}\}] \quad . \qquad \text{(E-16)}$$

Ordinarily $\lambda_j(t)$ is thought of as a deterministic but unknown function of t, i.e. time since start of the process. In an application to stratus forecasting, and to other weather phenomena, it may be desirable to allow a dependence of the basic hazard rate upon m, the duration or "age" of the current episode (stratus, or non-stratus as the case may be): $\lambda_j(m)$. This necessitates a specification, either parametric or non-parametric; the Cox procedure in Cox [1972] was to estimate $\lambda_j(m)$ non-parametrically.

In order to associate the Cox model explicitly with the logistic, adopt the attitude that $\lambda_j(m)$ is actually <u>random</u>, and is independently distributed from period to period, with a distribution characteristic of the state. In such a case we can do no better than to attempt to estimate the model

$$P_j(m,\underline{x}_t) = E(\exp[-\underline{\lambda}_j(m)\exp\{\underline{x}_t\underline{\beta}\}]) \;, \qquad \text{(E-17)}$$

where the expectation operator $E(\cdot)$ is over the distribution of the now-random hazard. To be quite specific, allow $\lambda_j(m)$ to have the Gamma distribution for $\alpha_j, \gamma_j > 0$,

$$P\{\underline{\lambda}_j(m) \leq x\} = \int_0^x e^{-\alpha_j y}\left(\frac{(\alpha_j y)^{\gamma_j - 1}}{\Gamma(\gamma_j)}\right)\gamma_j dy \;, \qquad \text{(E-18)}$$

where α_j and γ_j characterize the hazard variability when state j is in effect. Now for this distribution the expectation is explicitly in terms of the Laplace transform:

$$P_j(m,\underline{x}_t) = \int_0^{\infty} \exp[-y \exp\{\underline{x}_t\underline{\beta}\}] \frac{e^{-\alpha_j y}(\alpha_j y)^{\gamma_j}}{\Gamma(\gamma_j)} \gamma_j dy$$

$$= \left[\frac{\alpha_j}{\alpha_j + \exp\{\underline{x}_t\underline{\beta}\}}\right]^{\gamma_j}$$

$$= \left[\frac{1}{1 + \frac{1}{\alpha_j} e^{\underline{x}_t\underline{\beta}}}\right]^{\gamma_j} ; \qquad \text{(E-19)}$$

This is the probability of survival in state j for one more period (no change).

Now the probability of a change is, using the above randomizing model,

$$P\{Y_{t+1} \neq j | Y_t = j, x_t\} = 1 - \left[\frac{1}{1 + \frac{1}{\alpha_j} e^{\underline{x}_t\underline{\beta}}}\right]^{\gamma_j} \qquad \text{(E-20)}$$

and, in case $\gamma_j = 1$ (mixing by an exponential) we find

$$P\{Y_{t+1} \neq j | Y_t = j, \underline{X}_t = \underline{x}_t\} = \frac{\alpha_j^{-1} e^{\underline{x}_t\underline{\beta}}}{1 + \alpha_j^{-1} e^{\underline{x}_t\underline{\beta}}} \qquad \text{(E-21)}$$

which is precisely the logistic regression model. It is thus clear that the logistic regression model can arise from a plausible stochastic mechanism. Note that the derivation presents an alternative to the simple logistic model that incorporates one more parameter, thus possibly allowing for the better representation of a wider range of binary response data than by the classical logistic.

E.4 The Cox Survival Model with Stable-Law Random Hazard.

It is of interest to investigate other ways of introducing auxiliary randomness into the Cox proportional hazard survival model. This process considered here represents model parameter fluctuation from day to day (in the present application) that is not covered by the simple representation

$$P\{Y_{t+1}=j \mid Y_t=j, \underline{X}_t=\underline{x}_t\} = \exp[-\lambda \exp\{\underline{x}_t \underline{\beta}\}] \; ; \tag{E-22}$$

instead the form of the randomized model is obtained by inserting a term in the hazard:

$$P\{Y_{t+1}=j \mid Y_t=j, \underline{X}_t=\underline{x}_t, \varepsilon_t\} = \exp[-\lambda\varepsilon_t \exp\{\underline{x}_t \underline{\beta}\}] \; . \tag{E-23}$$

Now ε_t is not directly observable or estimable if, as is assumed, only one observation on a <u>probability</u> depending on each ε_t is available. Effectively one observes the marginal probability of $Y_{t+1} = j$, given $Y_t=j$ and values of the explanatory variables $\underline{X}_t$:

$$P\{Y_{t+1}=j \mid Y_t=j, \underline{X}_t=x_t\} = E_{\varepsilon_t}(\exp[-\lambda\varepsilon_t \exp\{\underline{x}_t \underline{\beta}\}]) \; . \tag{E-24}$$

Suppose now that ε_t obeys a <u>positive stable law</u> distribution (see Feller (1966), p. 170). In this case the Laplace transform of ε_t is always the form

$$E[e^{-s\varepsilon_t}] = e^{-(\alpha s)^{\gamma}} \; , \qquad 0 < \gamma < 1 \; . \tag{E-25}$$

Unfortunately, explicit formulas for the density of ε_t are generally not available; that for $\gamma = 1/2$ is an exception:

$$f_{\varepsilon_t}(x;\alpha,\tfrac{1}{2}) = \frac{1}{\sqrt{2\pi}(x/\alpha)^{3/2}} e^{-\alpha/2x} \; . \tag{E-26}$$

It follows generally and directly from (E-24) and (E-25) that if ε_t is positive stable the marginal probability of one-day survival is

$$P\{Y_{t+1}=j \mid Y_t=j, \underline{X}_t=\underline{x}_t\} = E_{\varepsilon_t}(\exp[-\lambda\varepsilon_t \exp\{\underline{x}_t\underline{\beta}\}])$$

$$= \exp[-(\lambda\alpha)^{\gamma} \exp\{\underline{x}_t\gamma\underline{\beta}\}] , \qquad \text{(E-27)}$$

once again exactly a Cox model (i.e. of the form (E-22)) but now with the parameters

$$\lambda' = (\lambda\alpha)^{\gamma} , \qquad \underline{\beta}' = \gamma\beta . \qquad \text{(E-28)}$$

Thus the particular Cox model discussed is completely insensitive to the type of hazard randomization introduced here. Notice that the effects of the explanatory variables or covariates, $\underline{x}_t$, as measured by the magnitudes of their coefficients ($\underline{\beta} \to \gamma\underline{\beta}$, $\gamma < 1$) , becomes progressively smaller as $\gamma \to 0$; the latter "shrinkage" tendency is associated with greater and greater "spread" of the ε_t distribution (here "spread" cannot be measured by variance, for the latter fails to exist). It follows that the predictive (in terms of explanatory variables) power of a Cox model could improve by reducing any tendency towards hazard randomization of the type exhibited, if such is possible.

Further work on randomized Cox models yielding binary time series will be reported elsewhere.

APPENDIX F

Spectral Analysis of Hourly Stratus Levels and Dew-Point Depression for July-September 1958.

The data for the height of the stratus level are hourly records, in units of hundreds of feet, of the height of the stratus layer. There are 2208 such observations. The data is integer valued with a minimum of three and a maximum of 999; 1410 of the observations are 999 which denotes the category of no stratus (infinite height); the next largest observational value is 888, of which there are 62; all the rest of the observations are less than or equal to 180.

Logarithms of the stratus heights were taken to reduce the range of the data. Figure (4) shows the ℓn (normalized periodogram) of the transformed data; (cf. Cox and Lewis (1966) pp. 99). If the data are uncorrelated and stationary then the values of the normalized periodogram will appear independent and have the unit exponential distribution. The line is at the 95% quantile for the maximum of 1104 independent unit exponentials. The largest peak occurs at 91. Other peaks occur at 1 and 276. The peak at 91 suggests that a 24 hour cycle may be present; the peak at 276 suggests an eight hour cycle. The peaks around 1 may be attributable to the dependence of the data. A least squares cyclic fit for the 24 and eight hour cycles was next carried out. The residuals from the fit were then whitened, using an AR2 process. Figure (5) shows the log (normalized periodogram) of the residuals following the cyclic fit and AR2 whitening. There are still

two values of the periodogram above the quantile line at 91 and 160. Figure (6) exhibits the cumulative periodogram of the residuals. If the residuals were uncorrelated and stationary, then the cumulative periodogram would have the same distribution as the order statistics of an independent sample of 1104 independent uniform random variables. The Kolmogorov-Smirnov statistic of goodness of fit is 1.12 (Theoretical 99% quantile is 1.628) and the Anderson-Darling statistic is 1.39 (theoretical 99% quantile is 3.857).

As a result of the above, we model the logarithm of hourly stratus heights as

$$\ell n\, L_t = (-1.55)\sin\left(\frac{2\pi t}{24}\right) - 0.322\,\cos\left(\frac{2\pi t}{24}\right)$$

$$(-0.202)\sin\left(\frac{2\pi t}{8}\right) - 0.300\,\cos\left(\frac{2\pi t}{8}\right) + A_t\ ;$$

$$A_t = 0.750A_{t-1} + 0.078A_{t-2} + E_t^{\ell}$$

where E_t^{ℓ} are stationary and uncorrelated random variables. Figure (7) shows the residuals E_t^{ℓ}.

A similar analysis was carried out on ℓn[dew point depression + 1] (LDPD). The data range from 0 to 9.21; the values have a discrete nature, but not as noticeably as that of the stratus levels. The ℓn-periodogram of LDPD is given in Figure 8. There are visible peaks at 92, 186 and 276, as well as near 1. The peaks at 92, 186, and 276 suggest 24 hr, 12 hr, and 8 hr cycles, respectively. A least-squares cyclic fit was made, and the residuals from the fit were once again whitened with an AR2 process.

Figure 9 gives the cumulative periodogram of the residuals with the Kolmogorov-Smirnov and Anderson-Darling statistics. Our model for LDPD is

$$\mathrm{LDPD}_t = 0.015 \sin\left(\frac{2\pi t}{24}\right) + 0.019 \cos\left(\frac{2\pi t}{24}\right)$$

$$+ 0.060 \sin\left(\frac{2\pi t}{12}\right) - 0.031 \cos\left(\frac{2\pi t}{12}\right)$$

$$+ 0.087 \sin\left(\frac{2\pi t}{8}\right) + 0.061 \cos\left(\frac{2\pi t}{8}\right)$$

$$+ B_t$$

$$B_t = .802\, B_{t-1} + .092\, B_{t-2} + E_t^d \,.$$

A graph of the residuals $\{E_t^d\}$ is presented in Figure 10. Note the two large residuals.

Next the residuals, $\{E_t^\ell\}$ of the ℓn (stratus height) level were regressed on $\{E_t^d\}$, the residuals of ℓn (dew point depression) using a least-squares procedure and the robust biweight procedure.

$$E_t^\ell = 0.0005 + 0.2712\, E_t^d \qquad \text{(Least Squares)}$$
$$(.022) \qquad (.076) \qquad \text{(Standard Errors)}$$

$$E_t^\ell = 0.0073 + 0.084\, E_t^d \qquad \text{(Biweight)}$$

If the two points corresponding to large LDPD residuals are deleted than the following values for regression coefficients are obtained

$$E_t^\ell = 0.009 + .3421\, E_t^d \qquad \text{(Least Squares)}$$
$$(.022)\ (.086) \qquad \text{(Standard Errors)}$$

$$E_t^d = 0.0083 + 0.1372\, E_t^d \qquad \text{(Biweight)}$$

The positive slope of the regression of the residuals suggest that the larger the dew point depression, the higher the stratus level. Since the regression was performed on the residuals of both series after detrending and whitening the relationship should not be strongly influenced by non-stationary and dependence effects in the marginal series. The small values of the fitted slopes suggest that the relationship is present, but not very strong. This relationship together with the box plots of Figure 1 provide evidence that dewpoint depression and the existence of stratus are indeed related.

The residuals were also examined for lagged relationships, eg. a relationship between E_t^{ℓ} and E_{t-1}^{d}. No relationships were evident.

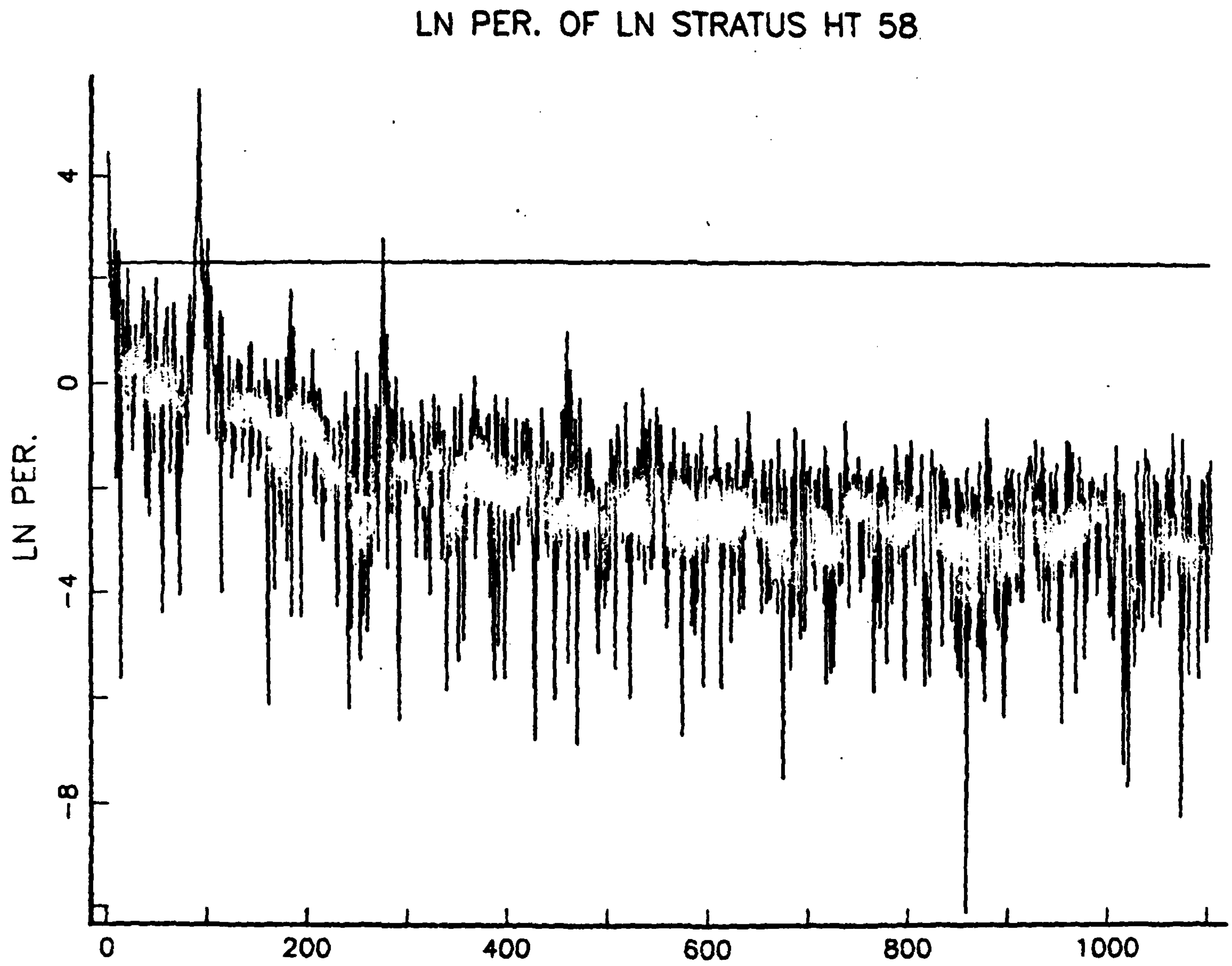


Figure 4

LN PER. OF AR2 RES OF 92,276 CY FIT LN STRATUS HT 58

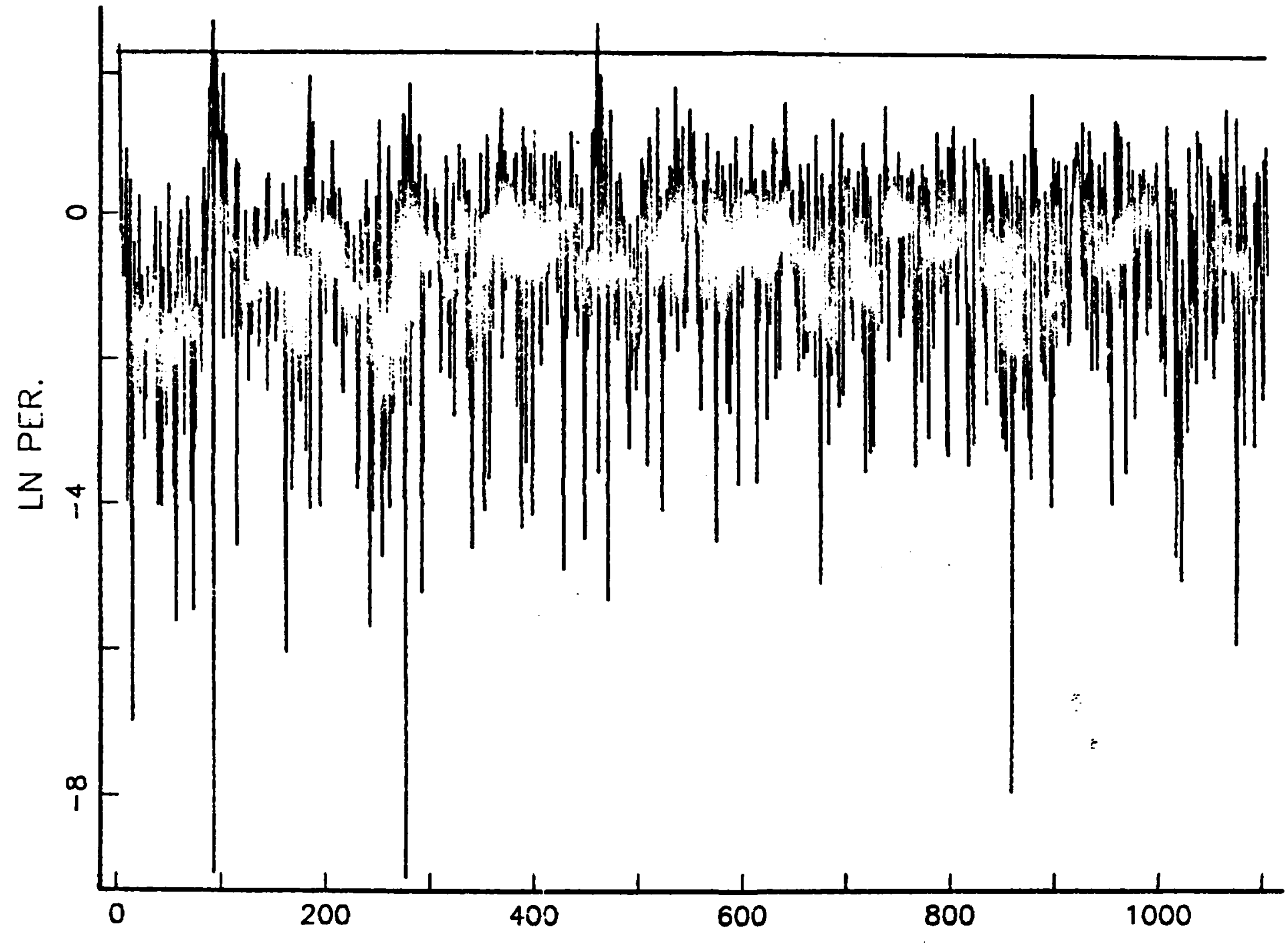


Figure 5

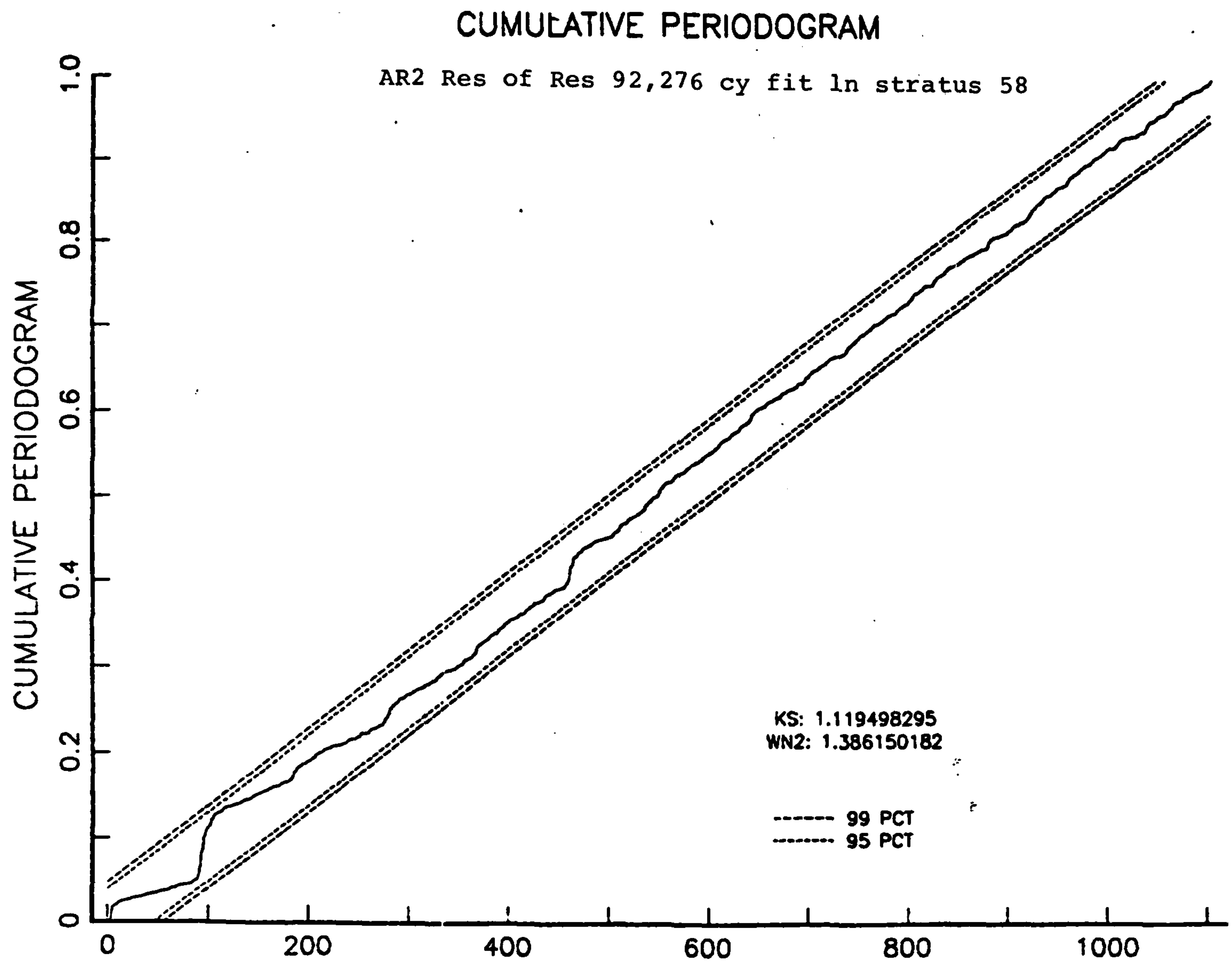

CUMULATIVE PERIODOGRAM
AR2 Res of Res 92,276 cy fit ln stratus 58
CUMULATIVE PERIODOGRAM
0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
KS: 1.119498295
WN2: 1.386150182
99 PCT
95 PCT

Figure 6

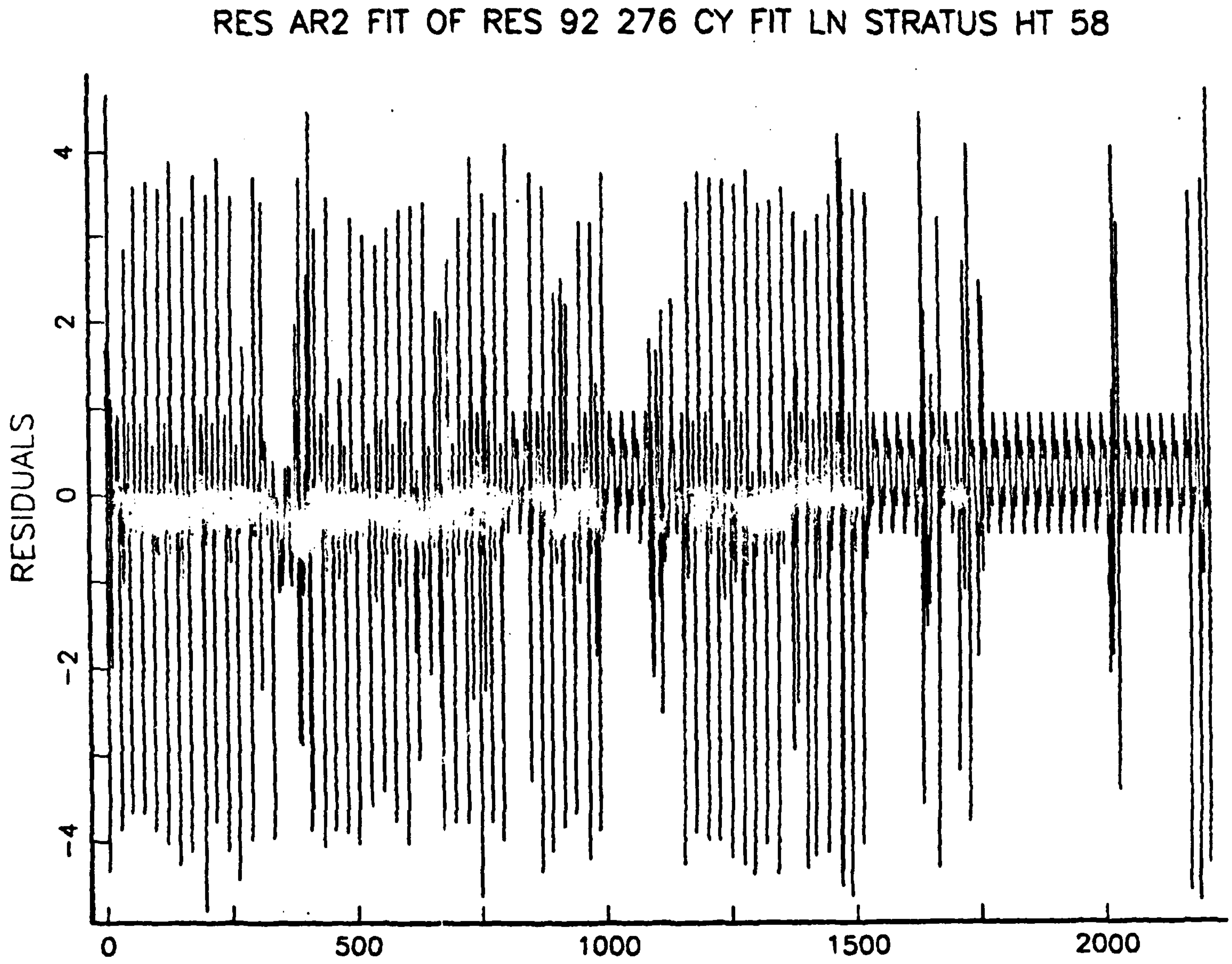

RES AR2 FIT OF RES 92 276 CY FIT LN STRATUS HT 58
RESIDUALS
4
2
0
-2
-4
0
500
1000
1500
2000

Figure 7

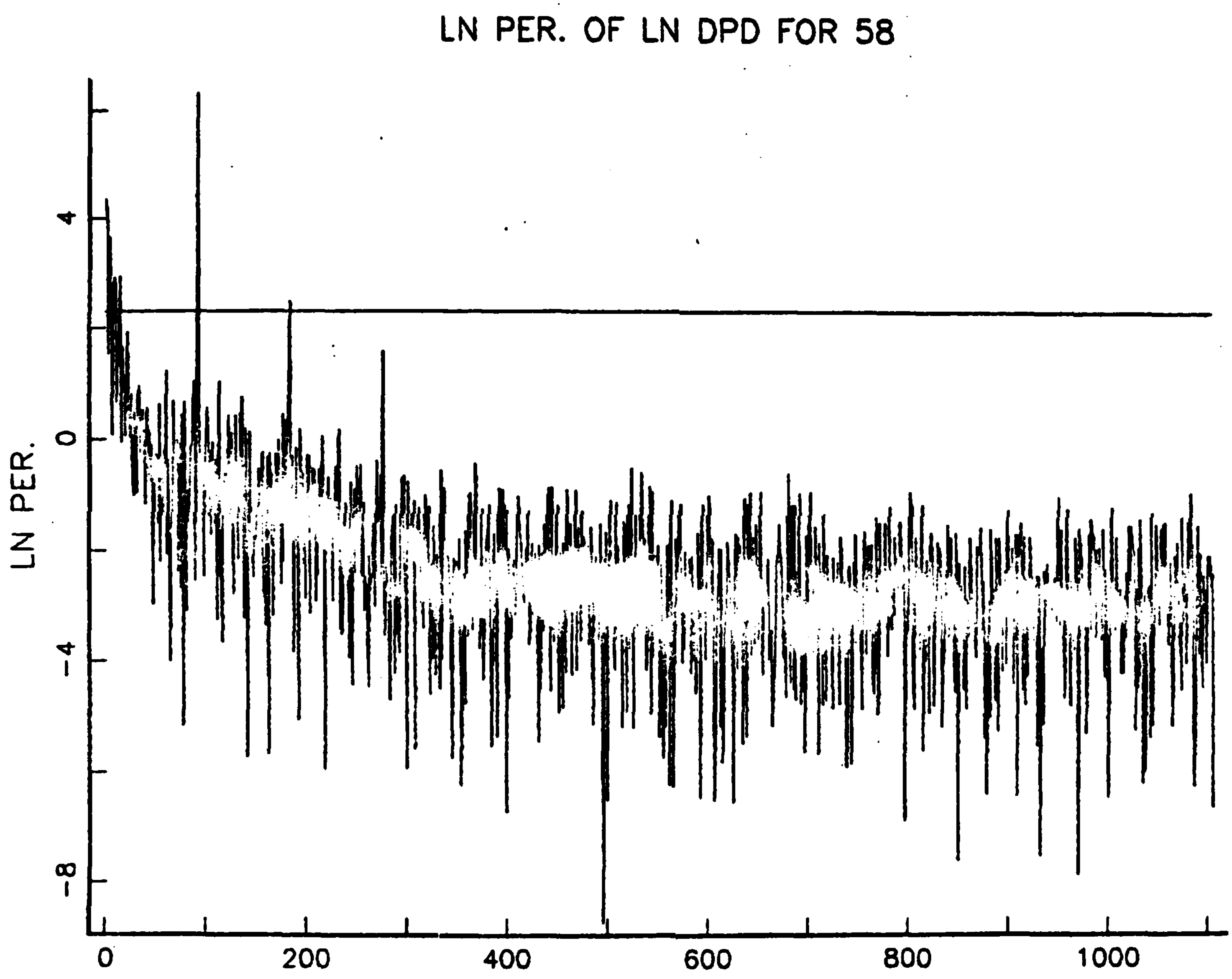
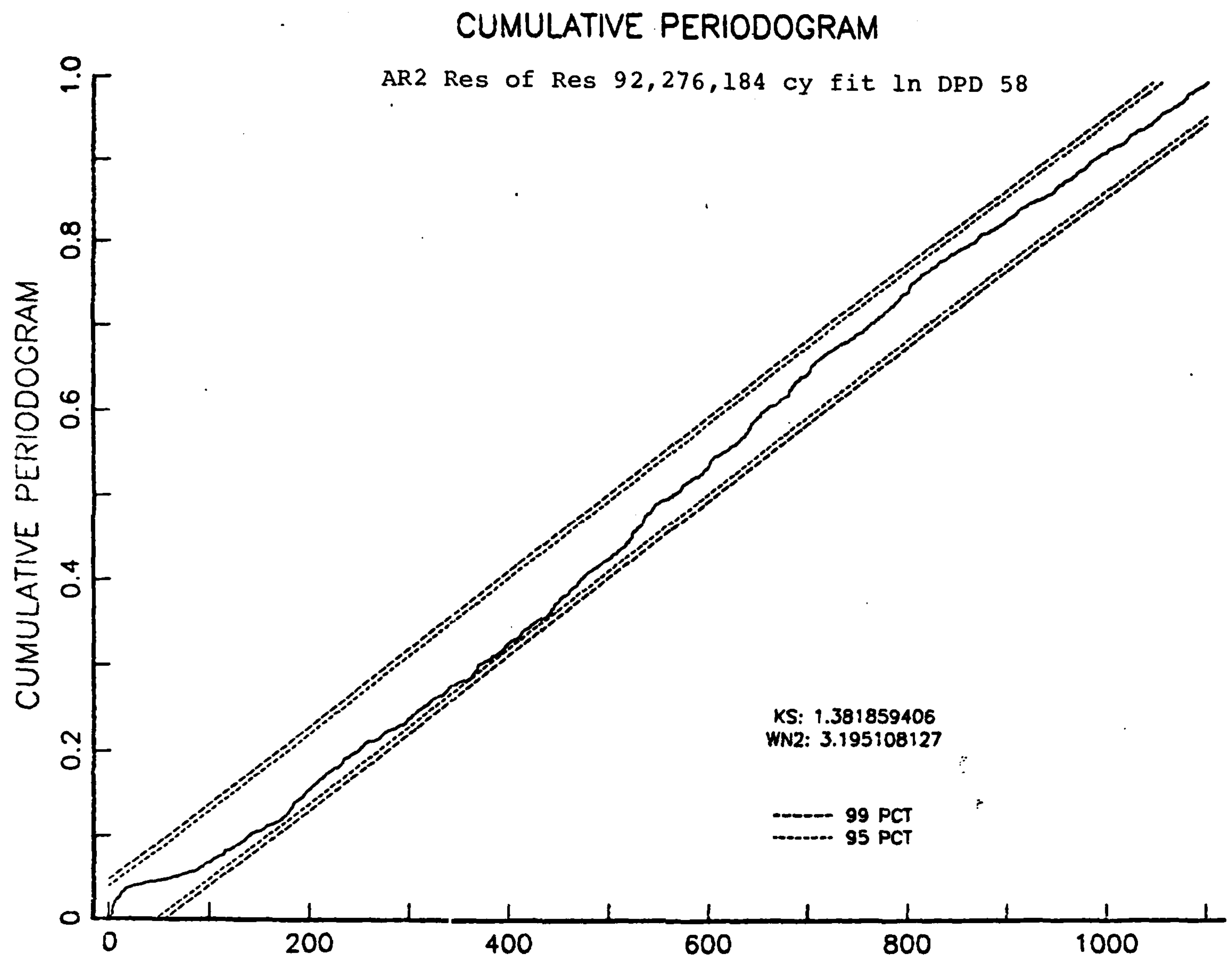


Figure 8



CUMULATIVE PERIODOGRAM
AR2 Res of Res 92,276,184 cy fit ln DPD 58
CUMULATIVE PERIODOGRAM
0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
KS: 1.381859406
WN2: 3.195108127
99 PCT
95 PCT

Figure 9

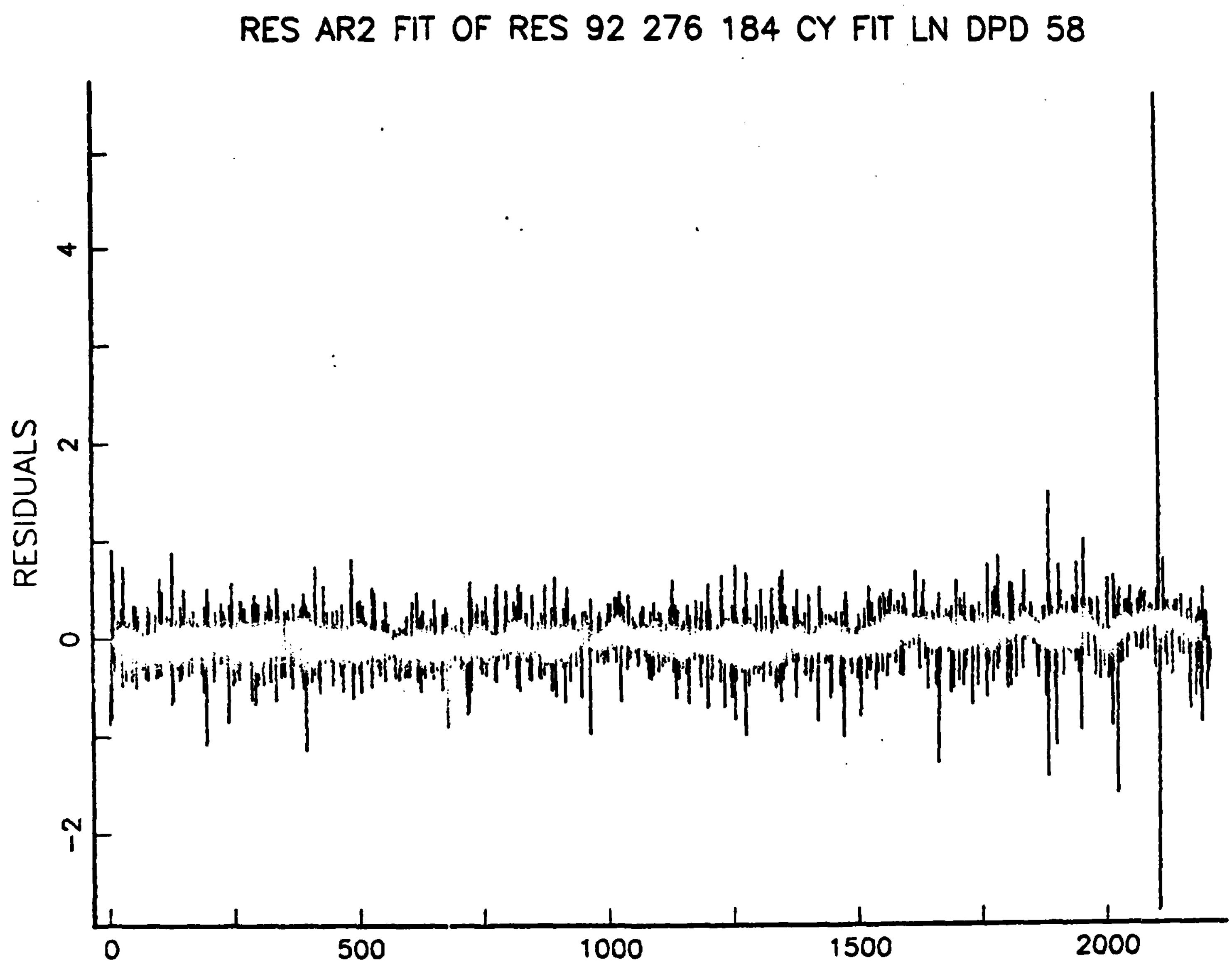

RES AR2 FIT OF RES 92 276 184 CY FIT LN DPD 58
RESIDUALS
4
2
0
-2
0
500
1000
1500
2000

Figure 10

APPENDIX G

Threat Scores

In this section we discuss the asymptotic distribution of a threat score.

Consider an event that either does or does not occur on day n, $n = 1,\ldots,N$.

Let

$$Y_n = \begin{cases} 1 & \text{if event occurs on day } n \text{;} \\ 0 & \text{if event does not occur on day } n \text{.} \end{cases}$$

Let

$$X_n = \begin{cases} 1 & \text{if the prediction is made that the event occurs on day } n \text{;} \\ 0 & \text{if the prediction is made that the event does not occur on day } n \text{.} \end{cases}$$

Let

$$S_0 = \sum_{n=1}^{N} (1-Y_n)(1-X_n)$$

be the number of correct predictions of the event not occurring;

$$S_1 = \sum_{n=1}^{N} Y_n X_n ,$$

be the number of correct predictions of the event occurring;

$$F_0 = \sum_{n=1}^{N} (1-Y_n) X_n ,$$

be the number of incorrect predictions when no event occurs;

$$F_1 = \sum_{n=1}^{N} Y_n(1-X_n) ,$$

the number of incorrect predictions when the event occurs.

The threat score for predicting that the event occurs is

$$T = \frac{S_1}{S_1 + F_0 + F_1} \qquad \text{(G-1)}$$

Equations (A-2), (A-3), and (A-4) give threat scores for predicting changes from no stratus to stratus, changes from stratus to no stratus, and all changes respectively.

Note that

$$T = \frac{\frac{S_1}{N}}{1-\frac{S_0}{N}} \equiv \left(f\ \frac{S_1}{N}, \frac{S_0}{N}\right) \qquad \text{(G-2)}$$

where $f(x_1,x_2) = \frac{x_1}{1-x_2}$

If there is perfect prediction, then $S_0 + S_1 = N$ and $T=1$. If $S_1=0$ then $T=0$. In the case of predicting changes from no stratus to stratus, the threat score would be 0 if prediction of stratus is done using only persistence.

Assume (S_0, F_0, S_1, F_1) has a multinomial distribution with parameters N, γ_{00}, γ_{01}, γ_{11}, γ_{10}. Asymptotically as $N \to \infty$, $\left(\frac{S_0}{N}, \frac{F_0}{N}, \frac{S_1}{N}, \frac{F_1}{N}\right)$ has a normal distribution with mean $(\gamma_{00}, \gamma_{01}, \gamma_{11}, \gamma_{10})$ and covariance matrix $\frac{1}{N} \sum$ where $\sum$ equals

$$\begin{vmatrix} \gamma_{00}(1-\gamma_{00}) & -\gamma_{00}\gamma_{01} & -\gamma_{00}\gamma_{11} & -\gamma_{00}\gamma_{10} \\ -\gamma_{00}\gamma_{01} & \gamma_{01}(1-\gamma_{01}) & -\gamma_{01}\gamma_{11} & -\gamma_{01}\gamma_{10} \\ -\gamma_{00}\gamma_{11} & -\gamma_{01}\gamma_{11} & \gamma_{11}(1-\gamma_{11}) & -\gamma_{11}\gamma_{10} \\ -\gamma_{00}\gamma_{10} & -\gamma_{01}\gamma_{10} & -\gamma_{11}\gamma_{10} & \gamma_{10}(1-\gamma_{10}) \end{vmatrix},$$

(cf. Bishop et al. (1975)).

A Taylor expansion of f in (G-2) yields

$$T = \frac{\gamma_{11}}{1-\gamma_{00}} + \frac{1}{1-\gamma_{00}}\left(\frac{S_0}{N} - \gamma_{00}\right) - \frac{\gamma_{11}}{1-\gamma_{00}{}^2}\left(\frac{S_1}{N} - \gamma_{11}\right)$$

$$+ o\left[\max\left(\frac{S_0}{N} - \gamma_{00}\right), \left(\frac{S_1}{N} - \gamma_{11}\right)\right].$$

It follows from an application of the multidimensional δ - method (cf. Theorem 14.6-2 of Bishop et al.) that as $N \to \infty$, T has an asymptotic normal distribution with mean $\frac{\gamma_{11}}{1-\gamma_{00}}$ and variance $\frac{1}{N}\sigma^2$ where

$$\sigma^2 = \gamma_{11}\frac{1-\gamma_{11}-\gamma_{00}}{(1-\gamma_{00})^3}$$

If γ_{00} is fixed, then σ^2 has a maximum at $\gamma_{11} = \frac{1-\gamma_{00}}{2}$ at which it has the value $\frac{1}{2(1-\gamma_{00})}$.

At $\gamma_{11} = 0$ and $\gamma_{11} = 1 - \gamma_{00}$, $\sigma^2 = 0$.

Another application of the δ-method shows that the transformed threat score $\arcsin\sqrt{T}$ has an asymptotic normal distribution with mean $\arcsin\sqrt{\frac{\gamma_{11}}{1-\gamma_{00}}}$ and variance

$$\frac{1}{N}\,\frac{1}{4}\,\frac{1}{(1-\gamma_{00})}\ .$$

Thus, if γ_{00} is fixed, then the transformed threat score $\arcsin\sqrt{T}$ has a variance which does not depend on γ_{11}. However, both the threat score, T, and $\arcsin\sqrt{T}$ can have large variance if γ_{00} is close to 1 (which will often be the case).

ACKNOWLEDGEMENTS

The authors wish to thank W. Sweet of NEPERF for providing the data and to Dr. A. Weinstein and to P. Lowe for useful comments; and to L. Uribe for his programming assistance. Figures 1-10 were produced by an experimental APL package GRAFSTAT which the Naval Postgraduate School is using under a test agreement with IBM Watson Research Center, Yorktown Heights, NY. We are grateful to Professor P.A.W. Lewis of the Naval Postgraduate School, and to Dr. P.D. Welch and Dr. P. Heidelberger of IBM for making this useful data-analytic implement available to us. The research of the authors was partially supported by the Naval Environmental Research Prediction Facility and by the Probability and Statistics Program of the Office of Naval Research.

References

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W., Discrete Multivariate Analysis: Theory and Practice. The M.I.T. Press. Cambridge, Massachusetts, and London, England. 1975.

Brelsford, W.M., and Jones, R.H., Estimating probabilities. Monthly Weather Review. Vol. 95. (1967). pp. 570-576.

Campbell, N.A., Robust procedures in multivariate analysis II. Robust canonical variate analysis. Applied Statistics. Vol. 31, No. 1. (1982) pp. 1-8.

Copas, J.B., Regression, prediction, and shrinkage. To appear, J. Royal Statist. Soc. B.

Cox, D.R., Analysis of Binary Data. Chapman and Hall, London. 1970.

Cox, D.R., Regression models of life tables (with discussion) J. R. Statist. Soc. B 33 (1972) pp. 187-220.

Cox, D.R., and Lewis. P.A.W., The Statistical Analysis of Series of Events. Methuen and Co., Ltd., London. 1966.

Feller, W., An Introduction to Probability Theory and Its Applications, II. John Wiley and Sons. 1966.

Gabriel, K.R., and Pun, F.C., Binary prediction of weather events with several predictors. Sixth Conference on Probability and Statistics in Atmospheric Sciences, American Meteorological Society, Boston, Mass. (1979) pp. 248-253.

Gilhousen, D.B., Testing the logit model for probability of precipitation forecasting. Sixth Conference on Probability and Statistics in Atmospheric Sciences, American Meteorological Soc., Boston, Mass. (1979) pp. 46-48.

Gnanadesikan, R., Methods for Statistical Data Analysis of Multivariate Data. John Wiley and Sons, New York. 1977.

Kolata, G., Computer graphics comes to statistics. Science 217 (Sept. 3, 1982). pp. 919-920.

Landwehr, J.M., Pregibon, D., and Shoemaker, A.C., Graphical methods for assessing logistic regression models. Bell Laboratories Technical Report. To appear in the Journal of the American Statistical Association.

Mosteller, F., and Tukey, J.W., Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley Publishing Company. Reading, Massachusetts. 1977.

Pregibon, D., Logistic regression diagnostics. Ann. Stat. 9 (1983). pp. 705-724.

__________. Reistant fits for some commonly used logistic models with medical applications. Biometrics 38 (1982) pp. 485-498.

DISTRIBUTION LIST

	NO. OF COPIES
Defense Technical Information Center Cameron Station Alexandria, VA 22314	2
Library Code 0142 Naval Postgraduate School	2
Dean of Research Code 012A Naval Postgraduate School Monterey, CA 93940	1
Library Code 55 Naval Postgraduate School Monterey, CA 93940	1
Professor M. L. Abdel-Hameed Department of Mathematics University of North Carolina Charlotte, NC 28223	1
Dr. G. P. Alldredge Department of Physics The University of Missouri Columbia, MO 65211	1
Professor F. J. Anscombe Department of Statistics Yale University, Box 2179 New Haven, CT 06520	1
Dr. Barbara Bailar Associate Director Statistical Standards Bureau of Census Washington, DC 20024	1
Mr. C. M. Bennett Code 741 Naval Coastal Systems Laboratory Panama City, FL 32401	1
Dr. Derrill J. Eordelon Code 21 Naval Underwater Systems Center Newport, RI 02840	1

	NO. OF COPIES
Dr. David Brillinger Statistics Department University of California Berkeley, CA 94720	1
Dr. R. W. Butterworth Systems Exploration 1340 Munras Avenue Monterey, CA 93490	1
Dr. D. R. Cox Department of Mathematics Imperial College London SW7 ENGLAND	1
Dr. D. F. Daley Statistics Department (IAS) Australian National University Canberra A.C.T. 2606 AUSTRALIA	1
Mr. DeSavage Naval Surface Weapons Center Silver Springs, MD 20910	1
Professor C. Derman Dept. of Civil Eng. & Mech. Engineering Columbia University New York, NY 10027	1
Dr. Guy Fayolle I.N.R.I.A. Dom de Voluceau-Rocquencourt 78150 Le Chesnay Cedex FRANCE	1
Dr. M. J. Fischer Defense Communications Agency 1860 Wiehle Avenue Reston, VA 22070	1
Professor George S. Fishman Cur. in OR & Systems Analysis University of North Carolina Chapel Hill, NC 20742	1
Dr. R. Gnanadesikan Bell Telephone Lab Murray Hill, NJ 07733	1

NO. OF COPIES

Professor Bernard Harris Department of Statistics University of Wisconsin 610 Walnut Street Madison, WI 53706	1
Dr. Gerhard Heiche Naval Air Systems Command (NAIR 03) Jefferson Plaza, No. 1 Arlington, VA 20360	1
Professor L. H. Herbach Department of Mathematics Polytechnic Institute of N.Y. Brooklyn, NY 11201	1
Professor W. M. Hinich University of Texas Austin, TX 78712	1
P. Heidelberger IBM Research Laboratory Yorktown Heights New York, NY 10598	1
W. D. Hibler, III Geophysical Fluid Dynamics Princeton University Princeton, NJ 08540	1
Professor D. L. Iglehart Department of Operations Research Stanford University Stanford, CA 94350	1
Dr. D. Vere Jones Department of Mathematics Victoria University of Wellington P. O. Box 196 Wellington NEW ZELAND	1
Professor J. B. Kadane Department of Statistics Carnegie-Mellon Pittsburgh, PA 15212	1

	NO. OF COPIES
Professor Guy Latouche University Libre Bruxelles C.P. 212 Blvd De Triomphe B-1050 Bruxelles BELGIUM	1
Dr. Richard Lau Offic of Naval Research Branch Office 1030 East Green Street Pasadena, CA 91101	1
A. J. Laurance Dept. of Mathematics Statistics University of Birmingham P. O. Box 363 Birmingham B15 2TT ENGLAND	1
Dr. John Copas Dept. of Mathematics Statistics University of Birmingham P. O. Box 363 Birmingham B15 2TT ENGLAND	1
Professor M. Leadbetter Department of Statistics University of North Carolina Chapel Hill, NC 27514	1
Mr. Dan Leonard Code 8105 Naval Ocean Systems Center San Diego, CA	1
M. Lepparanta Winter Navagation Res. Bd. Helsinki FINLAND	1
J. Lehoczky Department of Statistics Carnegie-Mellon University Pittsburgh, PA 15213	1
Library Naval Ocean Systems Center San Diego, CA 92132	1

NO. OF COPIES

Library 1
Code 1424
Naval Postgraduate School
Monterey, CA 93943

Dr. J. Maar (R51) 1
National Security Agency
Fort Meade, MD 20755

Bob Marcello 1
Canada Marine Engineering
Calgary
CANADA

Dr. M. McPhee 1
Chair of Arctic Marine Science
Oceanography Department
Naval Postgraduate School
Monterey, CA 93943

Dr. M. Mazumdar 1
Dept. of Industrial Engineering
University of Pittsburgh
Oakland
Pittsburgh, PA 15235

Professor Rupert G. Miller, Jr. 1
Statistics Department
Sequoia Hall
Stanford University
Stanford, CA 94305

National Science Foundation 1
Mathematical Sciences Section
1800 G Street, NW
Washington, DC 20550

Naval Research Laboratory 1
Technical Information Section
Washington, DC 20375

Professor Gordon Newell 1
Dept. of Civil Engineering
University of California
Berkeley, CA 94720

Dr. David Oakes 1
TUO Centenary Inst. of Occ. Health
London School of Hygiene/Tropical Med.
Keppel St. (Gower St.)
London WO1 E7H1
ENGLAND

	NO. OF COPIES
Dr. Alan F. Petty Code 7930 Navy Research Laboratory Washington, DC 20375	1
E. M. Reimnitz Pacific-Arctic Branch-Marine Geology U. S. Geological Survey 345 Middlefield Rd., (MS99) Menlo Park, CA 94025	1
Prof. M. Rosenblatt Department of Mathematics University of California - San Diego La Jolla, CA 92093	1
Professor I. R. Savage Department of Statistics Yale University New Haven, CT 06520	1
Professor W. R. Schucany Department of Statistics Southern Methodist University Dallas, TX 75222	1
Professor D. C. Siegmund Department of Statistics Sequoia Hall Stanford University Stanford, CA 94305	1
Professor H. Solomon Department of Statistics Sequoia Hall Stanford University Stanford, CA 94305	1
Dr. Ed Wegman Statistics & Probability Program Code 411(SP) Office of Naval Research Arlington, VA 22217	1
Dr. Douglas de Priest Statistics & Probability Program Code 411(SP) Office of Naval Research Arlington, VA 22217	1

NO. OF COPIES

Dr. Marvin Moss Statistics & Probability Program Code 411(SP) Office of Naval Research Arlington, VA 22217	1
Technical Library Naval Ordnance Station Indian Head, MD 20640	1
Professor J. R. Thompson Dept. of Mathematical Science Rice University Houston, TX 77001	1
Professor J. W. Tukey Statistics Department Princeton University Princeton, NJ 08540	1
P. Wadhams Scott Polar Research Cambridge University Cambridge CB2 1ER ENGLAND	1
Daniel H. Wagner Station Square One Paoli, PA 19301	1
Dr. W. Weeks U.S. Army CR REL 72 Lyme Road Hanover, NH 03755	1
P. Welch IBM Research Laboratory Yorktown Heights, NY 10598	1
Pat Welsh Head, Polar Oceanography Branch Code 332 Naval Ocean Research & Dev. Activity NSTL Station Mississippi 39529	1
Dr. Roy Welsch Sloan School M.I.T. Cambridge, MA 02139	1

NO. OF COPIES

Dr. Morris DeGroot Statistics Department Carnegie-Mellon University Pittsburgh, PA 15235	1
Professor R. Renard Head, Meteorology Department Naval Postgraduate School Monterey, CA 93943	1
A. Weinstein Commanding Officer Naval Environmental Prediction Research Facility Monterey, CA 93943	1
Paul Lowe Naval Environmental Prediction Research Facility Monterey, CA 93943	1
Wayne Sweet Naval Environmental Prediction Research Facility Monterey, CA 93943	1
Dr. Colin Mallows Bell Telephone Laboratories Murray Hill, NJ 07974	1
Dr. D. Pregibon Bell Telephone Laboratories Murray Hill, NJ 07974	1
Dr. Jon Kettenring Bell Telephone Laboratories Murray Hill, NJ 07974	1
Professor Grace Woehba Department of Statistics University of Wisconsin 1210 W. Dayton St. Madison, WI 53706	1
Professor D. Gaver Code 55Gv Naval Postgraduate School Monterey, CA 93941	25
Professor P. Jacobs Code 55Jc Naval Postgraduate School Monterey, CA 93941	10